

DIAGNOSTIC AND THERAPEUTIC METHODS USING MOLECULES
DIFFERENTIALLY EXPRESSED IN CANCER CELLS

Cross-Reference to Related Applications

This application claims the benefit of U.S. Provisional Application No. 60/101,900, filed September 25, 1998, the entirety of which is incorporated herein by reference.

Field of the Invention

This invention relates to methods for predicting and influencing the behavior of cells 5 and tumors. In particular embodiments, the invention relates to methods in which a cell is examined for expression of a specified gene sequence to determine propensity for metastatic spread. In other embodiments, the invention relates to the inhibition of metastatic spread.

Background of the Invention

Breast, colon, and lung cancers represent the most common cancers. Despite the use 10 of a number of histochemical, genetic, and immunological markers, clinicians still have a difficult time predicting which tumors will metastasize to other organs using conventional methodologies. Some patients are in need of adjuvant therapy to prevent recurrence and metastasis and others are not. However, distinguishing between these subpopulations of patients is not straightforward. Thus, the course of treatment is not easily charted. There is a 15 need in the art for new markers for distinguishing between normal tissue and tumor tissue, and between tumors which will or have spread and those which are less likely to metastasize.

Summary of the Invention

The invention provides materials and methods for determining the metastatic potential 20 of a cell and for identifying cancerous cells by determining the presence or absence of one or more expression products of at least one gene that is differentially expressed between normal cells, nonmetastatic cells, cells of low metastatic potential, and cells of high metastatic potential.

In one embodiment, the invention features a method for assessing the metastatic 25 potential of a breast cell, a colon cell, or a lung cell by detecting expression of a differentially expressed gene in a test cell and comparing expression of the gene in a control cell, wherein the level of expression of the gene in the test cell relative to the level of expression in the

control cell is indicative of the metastatic potential of the test cell. In general, the differentially expressed gene comprises a sequence selected from the group consisting of SEQ ID NOS:1-37.

In another embodiment, the invention features a method for detecting a cancerous breast, colon, lung, or pancreas cell by detecting expression of a differentially expressed gene in a test cell and comparing expression of the gene in a control cell, wherein the level of expression of the gene in the test cell relative to the level of expression in the control cell is indicative of the metastatic potential of the test cell. In general, the differentially expressed gene comprises a sequence selected from the group consisting of SEQ ID NOS:1-37.

In other embodiments, the invention features inhibition of metastasis of a cancerous cell by expression of a gene that is overexpressed in cells of low metastatic potential, or by inhibiting expression of a gene overexpressed in cells of high metastatic potential.

These and other embodiments of the invention will be readily apparent upon reading the specification provided herein. The invention will now be described in more detail.

15 Detailed Description of the Invention

The present invention is based on the discovery of that specific genes identified herein are differentially expressed between normal cells, non-metastatic cancer cells, and metastatic cancer cells, which cells are obtained from different tissue types, including breast, lung and colon. This differential expression information can be exploited in diagnostics using diagnostic reagents specific for the expression products of the differentially expressed gene. This information can also be used in diagnostic and prognostic methods which will help clinicians in planning appropriate treatment regimens for cancers, including cancer of the breast, lung or colon. Identification of these differentially expressed polynucleotides also permits the formulation of diagnostic and therapeutic reagents and methods as further described below.

Diagnostic and Prognostic Methods

The invention provides a method for determining the metastatic potential of a cell by determining the presence or absence of an expression product of a gene that is preferentially expressed in normal cells and cells having low metastatic potential as compared to cells from highly metastatic cell lines. Other methods are used for determining the presence, absence, or relative levels of an expression product preferentially expressed in cells having high

metastatic potential. The presence, absence, or relative level of an expression product of the gene can be determined by, for example, examining RNA levels (*e.g.*, by directly detecting RNA, or by producing cDNA from the RNA and detecting levels of the relevant cDNAs) or detection of a polypeptide encoded by the differentially expressed gene.

- 5 In one exemplary embodiment, the method comprises determining the metastatic potential of a cell by detecting the relative expression level of a differentially expressed gene corresponding to a polynucleotide comprising a sequence of at least one of SEQ ID NOS:1-37, where the relative expression level is determined by comparing an expression level of the differentially expressed gene in a sample obtained from the cell to a level of expression of a
10 control gene (*e.g.*, a housekeeping gene or other gene unaffected by the cancerous state of the cell) and/or to other differentially expressed genes corresponding to a polynucleotide described herein. In another exemplary embodiment, the invention comprises detection of a cancerous cell, *e.g.*, a cell having tumor potential, a non-metastatic cell, a cell of low metastatic potential (*e.g.*, a cell having a low probability of progressing to a metastasis,
15 including nonmetastatic cells), a high metastatic potential cell, or a metastasized cell. In general, the cells analyzed according to the methods of the invention are obtained from a patient having, susceptible to (*e.g.*, having a family history or other risk factor), or suspected of having cancer. Examples of cancer include, but are not limited to, breast, colon and lung cancer.
20 The methods and other aspects of the invention are described below in more detail.

Polynucleotide Compositions

Polynucleotide compositions useful in the methods of the invention include, but are not necessarily limited to, polynucleotides having a sequence set forth in any one of SEQ ID NOS:1-37; polynucleotides obtained from the biological materials described herein or other biological sources (particularly human sources) by hybridization under stringent conditions (particularly conditions of high stringency); genes corresponding to the provided polynucleotides; variants of the provided polynucleotides and their corresponding genes, particularly those variants that retain a biological activity of the encoded gene product (*e.g.*, a
25 biological activity ascribed to a gene product corresponding to the provided polynucleotides as a result of the assignment of the gene product to a protein family(ies) and/or identification of a functional domain present in the gene product). Other nucleic acid compositions useful in and contemplated by the present invention will be readily apparent to one of ordinary skill
30

in the art when provided with the disclosure here. "Polynucleotide" and "nucleic acid" as used herein with reference to nucleic acids of the composition is not intended to be limiting as to the length or structure of the nucleic acid unless specifically indicated.

The invention features polynucleotides that are differentially expressed in human tissue, specifically human colon, breast, and/or lung tissue. Nucleic acid compositions of particular interest for use in the invention comprise a sequence set forth in any one of SEQ ID NOS:1-37 or an identifying sequence thereof. An "identifying sequence" is a contiguous sequence of residues at least about 10 nt to about 20 nt in length, usually at least about 50 nt to about 100 nt in length, that uniquely identifies a polynucleotide sequence, *e.g.*, exhibits less than 90%, usually less than about 80% to about 85% sequence identity to any contiguous nucleotide sequence of more than about 20 nt. Thus, the subject novel nucleic acid compositions include full length cDNAs or mRNAs that encompass an identifying sequence of contiguous nucleotides from any one of SEQ ID NOS: 1-37.

Polynucleotides useful and contemplated for use in the present invention also include polynucleotides having sequence similarity or sequence identity to the provided polynucleotides. Nucleic acids having sequence similarity are detected by hybridization under low stringency conditions, for example, at 50°C and 10XSSC (0.9 M saline/0.09 M sodium citrate) and remain bound when subjected to washing at 55°C in 1XSSC. Sequence identity can be determined by hybridization under stringent conditions, for example, at 50°C or higher and 0.1XSSC (9 mM saline/0.9 mM sodium citrate). Hybridization methods and conditions are well known in the art, see, *e.g.*, USPN 5,707,829. Nucleic acids that are substantially identical to the provided polynucleotide sequences, *e.g.* allelic variants, genetically altered versions of the gene, *etc.*, bind to the provided polynucleotide sequences (SEQ ID NOS:1-37) under stringent hybridization conditions. By using probes, particularly labeled probes of DNA sequences, one can isolate homologous or related genes. The source of homologous genes can be any species, *e.g.* primate species, particularly human; rodents, such as rats and mice; canines, felines, bovines, ovines, equines, yeast, nematodes, *etc.*

Preferably, hybridization is performed using at least 15 contiguous nucleotides (nt) of at least one of SEQ ID NOS:1-37. That is, when at least 15 contiguous nt of one of the disclosed SEQ ID NOS. is used as a probe, the probe will preferentially hybridize with a nucleic acid comprising the complementary sequence, allowing the identification and retrieval of the nucleic acids that uniquely hybridize to the selected probe. Probes from more than one SEQ ID NO. can hybridize with the same nucleic acid if the cDNA from which they

were derived corresponds to one mRNA. Probes of more than 15 nt can be used, e.g., probes of from about 18 nt to about 100 nt, but 15 nt represents sufficient sequence for unique identification.

Polynucleotides useful in the invention also include naturally occurring variants of the provided nucleotide sequences (e.g., degenerate variants, allelic variants, etc.). Variants of the polynucleotides of the invention are identified by hybridization of putative variants with nucleotide sequences disclosed herein, preferably by hybridization under stringent conditions. For example, by using appropriate wash conditions, variants of the polynucleotides of the invention can be identified where the allelic variant exhibits at most about 25-30% base pair (bp) mismatches relative to the selected polynucleotide probe. In general, allelic variants contain 15-25% bp mismatches, and can contain as little as even 5-15%, or 2-5%, or 1-2% bp mismatches, as well as a single bp mismatch.

The invention also encompasses use of homologs corresponding to the polynucleotides of SEQ ID NOS:1-37, where the source of homologous genes can be any mammalian species, e.g., primate species, particularly human; rodents, such as rats; canines, felines, bovines, ovines, equines, yeast, nematodes, etc. Between mammalian species, e.g., human and mouse, homologs generally have substantial sequence similarity, e.g., at least 75% sequence identity, usually at least 90%, more usually at least 95% between nucleotide sequences. Sequence similarity is calculated based on a reference sequence, which may be a subset of a larger sequence, such as a conserved motif, coding region, flanking region, etc. A reference sequence will usually be at least about 18 contiguous nt long, more usually at least about 30 nt long, and may extend to the complete sequence that is being compared.

Algorithms for sequence analysis are known in the art, such as gapped BLAST, described in Altschul, et al. *Nucleic Acids Res.* (1997) 25:3389-3402.

In general, variant polynucleotides of the invention have a sequence identity greater than at least about 65%, preferably at least about 75%, more preferably at least about 85%, and can be greater than at least about 90% or more as determined by the Smith-Waterman homology search algorithm as implemented in MPSRCH program (Oxford Molecular). For the purposes of this invention, a preferred method of calculating percent identity is the Smith-Waterman algorithm, using the following. Global DNA sequence identity must be greater than 65% as determined by the Smith-Waterman homology search algorithm as implemented in Smith-Waterman (Time Logic) program using an affine gap search with the following search parameters: gap open penalty, 12; and gap extension penalty, 1.

Fragments of the provided polynucleotides can be used in the invention, particularly fragments that encode a unique identifier of a differentially expressed gene of interest, *etc.*

The term "cDNA" as used herein is intended to include all nucleic acids that share the arrangement of sequence elements found in native mature mRNA species, where sequence

- 5 elements are exons and 3' and 5' non-coding regions. Normally mRNA species have contiguous exons, with the intervening introns, when present, being removed by nuclear RNA splicing, to create a continuous open reading frame encoding a polypeptide of the invention.

A genomic sequence comprises the nucleic acid present between the initiation codon

- 10 and the stop codon, as defined in the listed sequences, including all of the introns that are normally present in a native chromosome. It can further include the 3' and 5' untranslated regions found in the mature mRNA. It can further include specific transcriptional and translational regulatory sequences, such as promoters, enhancers, *etc.*, including about 1 kb, but possibly more, of flanking genomic DNA at either the 5' and 3' end of the transcribed
15 region. The genomic DNA can be isolated as a fragment of 100 kbp or smaller; and substantially free of flanking chromosomal sequence. The genomic DNA flanking the coding region, either 3' and 5', or internal regulatory sequences as sometimes found in introns, contains sequences required for proper tissue, stage-specific, or disease-state specific expression.

- 20 The nucleic acid compositions of the subject invention can encode all or a part of the polypeptides encoded by the gene corresponding to the provided polynucleotides. Double or single stranded fragments can be obtained from the DNA sequence by chemically synthesizing oligonucleotides in accordance with conventional methods, by restriction enzyme digestion, by PCR amplification, *etc.* Isolated polynucleotides and polynucleotide
25 fragments of the invention comprise at least about 10, about 15, about 20, about 35, about 50, about 100, about 150 to about 200, about 250 to about 300, or about 350 contiguous nt selected from the polynucleotide sequences as shown in SEQ ID NOS:1-37. For the most part, fragments will be of at least 15 nt, usually at least 18 nt or 25 nt, and up to at least about 50 contiguous nt in length or more. In a preferred embodiment, the polynucleotide molecules
30 comprise a contiguous sequence of at least 12 nt selected from the group consisting of the polynucleotides shown in SEQ ID NOS:1-37.

Probes specific to the genes corresponding to the provided polynucleotides can be generated using the polynucleotide sequences disclosed in SEQ ID NOS:1-37. The probes

are preferably at least about a 12, 15, 16, 18, 20, 22, 24, or 25 nt fragment of a corresponding contiguous sequence of SEQ ID NOS:1-37, and can be less than 2, 1, 0.5, 0.1, or 0.05 kb in length. The probes can be synthesized chemically or can be generated from longer polynucleotides using restriction enzymes. The probes can be labeled, for example, with a 5 radioactive, biotinylated, or fluorescent tag. Preferably, probes are designed based upon an identifying sequence of a polynucleotide of one of SEQ ID NOS:1-37. More preferably, probes are designed based on a contiguous sequence of one of the subject polynucleotides that remain unmasked following application of a masking program for masking low complexity (e.g., XBLAST) to the sequence., i.e., one would select an unmasked region, as 10 indicated by the polynucleotides outside the poly-n stretches of the masked sequence produced by the masking program.

The polynucleotides can be isolated and obtained in substantial purity, generally as other than an intact chromosome. Usually, the polynucleotides, either as DNA or RNA, will be obtained substantially free of other naturally-occurring nucleic acid sequences, generally 15 being at least about 50%, usually at least about 90% pure and are typically “recombinant”, e.g., flanked by one or more nucleotides with which it is not normally associated on a naturally occurring chromosome.

The polynucleotides of the invention can be provided as a linear molecule or within a circular molecule, and can be provided within autonomously replicating molecules (vectors) 20 or within molecules without replication sequences. Expression of the polynucleotides can be regulated by their own or by other regulatory sequences known in the art. The polynucleotides of the invention can be introduced into suitable host cells using a variety of techniques available in the art, such as transferrin polycation-mediated DNA transfer, transfection with naked or encapsulated nucleic acids, liposome-mediated DNA transfer, 25 intracellular transportation of DNA-coated latex beads, protoplast fusion, viral infection, electroporation, gene gun, calcium phosphate-mediated transfection, and the like.

The subject nucleic acid compositions can be used to, for example, produce polypeptides, as probes for the detection of mRNA of the invention in biological samples (e.g., extracts of human cells) to generate additional copies of the polynucleotides, to 30 generate ribozymes or antisense oligonucleotides, and as single stranded DNA probes or as triple-strand forming oligonucleotides. The probes described herein can be used to, for example, detect in a sample the presence, absence, and/or relative levels of gene products

corresponding to the polynucleotide sequences as shown in SEQ ID NOS:1-37 or variants thereof. These and other uses are described in more detail below.

Use of Polynucleotides to Obtain Full-Length cDNA, Gene, and Promoter Region

Full-length cDNA molecules comprising the disclosed polynucleotides are obtained

- 5 as follows. A polynucleotide having a sequence of one of SEQ ID NOS:1-37, or a portion thereof comprising at least 12, 15, 18, or 20 nt, is used as a hybridization probe to detect hybridizing members of a cDNA library using probe design methods, cloning methods, and clone selection techniques such as those described in USPN 5,654,173. Libraries of cDNA are made from selected tissues, such as normal or tumor tissue, or from tissues of a mammal
- 10 treated with, for example, a pharmaceutical agent. Preferably, the tissue is the same as the tissue from which the polynucleotides of the invention were isolated, as both the polynucleotides described herein and the cDNA represent expressed genes. Most preferably, the cDNA library is made from the biological material described herein in the Examples. The choice of cell type for library construction can be made after the identity of the protein
- 15 encoded by the gene corresponding to the polynucleotide of the invention is known. This will indicate which tissue and cell types are likely to express the related gene, and thus represent a suitable source for the mRNA for generating the cDNA. Where the provided polynucleotides are isolated from cDNA libraries, the libraries are prepared from mRNA of human colon cells, more preferably, human colon cancer cells, even more preferably, from a
- 20 highly metastatic colon cell, Km12L4-A.

Techniques for producing and probing nucleic acid sequence libraries are described, for example, in Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, 2nd Ed., (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY. The cDNA can be prepared by using primers based on sequence from SEQ ID NOS:1-37. In one embodiment, the cDNA library can be made from only poly-adenylated mRNA. Thus, poly-T primers can be used to prepare cDNA from the mRNA.

Members of the library that are larger than the provided polynucleotides, and preferably that encompass the complete coding sequence of the native message, are obtained. In order to confirm that the entire cDNA has been obtained, RNA protection experiments are performed as follows. Hybridization of a full-length cDNA to an mRNA will protect the RNA from RNase degradation. If the cDNA is not full length, then the portions of the mRNA that are not hybridized will be subject to RNase degradation. This is assayed, as is known in the art, by changes in electrophoretic mobility on polyacrylamide gels, or by

detection of released monoribonucleotides. Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual, 2nd Ed.*, (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY. In order to obtain additional sequences 5' to the end of a partial cDNA, 5' RACE (*PCR Protocols: A Guide to Methods and Applications*, (1990) Academic Press, Inc.) can be

5 performed.

Genomic DNA is isolated using the provided polynucleotides in a manner similar to the isolation of full-length cDNAs. Briefly, the provided polynucleotides, or portions thereof, are used as probes to libraries of genomic DNA. Preferably, the library is obtained from the cell type that was used to generate the polynucleotides of the invention, but this is not

10 essential. Most preferably, the genomic DNA is obtained from the biological material described herein in the Examples. Such libraries can be in vectors suitable for carrying large segments of a genome, such as P1 or YAC, as described in detail in Sambrook *et al.*, 9.4-9.30. In addition, genomic sequences can be isolated from human BAC libraries, which are commercially available from Research Genetics, Inc., Huntsville, Alabama, USA, for

15 example. In order to obtain additional 5' or 3' sequences, chromosome walking is performed, as described in Sambrook *et al.*, such that adjacent and overlapping fragments of genomic DNA are isolated. These are mapped and pieced together, as is known in the art, using restriction digestion enzymes and DNA ligase.

Using the polynucleotide sequences of the invention, corresponding full-length genes

20 can be isolated using both classical and PCR methods to construct and probe cDNA libraries. Using either method, Northern blots, preferably, are performed on a number of cell types to determine which cell lines express the gene of interest at the highest level. Classical methods of constructing cDNA libraries are taught in Sambrook *et al.*, supra. With these methods, cDNA can be produced from mRNA and inserted into viral or expression vectors. Typically,

25 libraries of mRNA comprising poly(A) tails can be produced with poly(T) primers. Similarly, cDNA libraries can be produced using the instant sequences as primers.

PCR methods are used to amplify the members of a cDNA library that comprise the desired insert. In this case, the desired insert will contain sequence from the full length cDNA that corresponds to the instant polynucleotides. Such PCR methods include gene

30 trapping and RACE methods. Gene trapping entails inserting a member of a cDNA library into a vector. The vector then is denatured to produce single stranded molecules. Next, a substrate-bound probe, such a biotinylated oligo, is used to trap cDNA inserts of interest. Biotinylated probes can be linked to an avidin-bound solid substrate. PCR methods can be

used to amplify the trapped cDNA. To trap sequences corresponding to the full length genes, the labeled probe sequence is based on the polynucleotide sequences of the invention.

Random primers or primers specific to the library vector can be used to amplify the trapped cDNA. Such gene trapping techniques are described in Gruber et al., WO 95/04745 and

- 5 Gruber et al., USPN 5,500,356. Kits are commercially available to perform gene trapping experiments from, for example, Life Technologies, Gaithersburg, Maryland, USA.

“Rapid amplification of cDNA ends,” or RACE, is a PCR method of amplifying cDNAs from a number of different RNAs. The cDNAs are ligated to an oligonucleotide linker, and amplified by PCR using two primers. One primer is based on sequence from the 10 instant polynucleotides, for which full length sequence is desired, and a second primer comprises sequence that hybridizes to the oligonucleotide linker to amplify the cDNA. A description of this methods is reported in WO 97/19110. In preferred embodiments of RACE, a common primer is designed to anneal to an arbitrary adaptor sequence ligated to cDNA ends (Apte and Siebert, *Biotechniques* (1993) 15:890-893; Edwards et al., *Nuc. Acids Res.* (1991) 19:5227-5232). When a single gene-specific RACE primer is paired with the common primer, preferential amplification of sequences between the single gene specific primer and the common primer occurs. Commercial cDNA pools modified for use in RACE 15 are available.

Another PCR-based method generates full-length cDNA library with anchored ends 20 without needing specific knowledge of the cDNA sequence. The method uses lock-docking primers (I-VI), where one primer, poly TV (I-III) locks over the polyA tail of eukaryotic mRNA producing first strand synthesis and a second primer, polyGH (IV-VI) locks onto the polyC tail added by terminal deoxynucleotidyl transferase (TdT)(see, e.g., WO 96/40998).

The promoter region of a gene generally is located 5' to the initiation site for RNA 25 polymerase II. Hundreds of promoter regions contain the “TATA” box, a sequence such as TATTA or TATAA, which is sensitive to mutations. The promoter region can be obtained by performing 5' RACE using a primer from the coding region of the gene. Alternatively, the cDNA can be used as a probe for the genomic sequence, and the region 5' to the coding 30 region is identified by “walking up.” If the gene is highly expressed or differentially expressed, the promoter from the gene can be of use in a regulatory construct for a heterologous gene.

Once the full-length cDNA or gene is obtained, DNA encoding variants can be prepared by site-directed mutagenesis, described in detail in Sambrook et al., 15.3-15.63. The

choice of codon or nucleotide to be replaced can be based on disclosure herein on optional changes in amino acids to achieve altered protein structure and/or function.

- As an alternative method to obtaining DNA or RNA from a biological material, nucleic acid comprising nucleotides having the sequence of one or more polynucleotides of
- 5 the invention can be synthesized. Thus, the invention encompasses nucleic acid molecules ranging in length from 15 nt (corresponding to at least 15 contiguous nt of one of SEQ ID NOS:1-37) up to a maximum length suitable for one or more biological manipulations, including replication and expression, of the nucleic acid molecule. The invention includes but is not limited to (a) nucleic acid having the size of a full gene, and comprising at least one
- 10 of SEQ ID NOS:1-37; (b) the nucleic acid of (a) also comprising at least one additional gene, operably linked to permit expression of a fusion protein; (c) an expression vector comprising (a) or (b); (d) a plasmid comprising (a) or (b); and (e) a recombinant viral particle comprising (a) or (b). Once provided with the polynucleotides disclosed herein, construction or preparation of (a) - (e) are well within the skill in the art.
- 15 The sequence of a nucleic acid comprising at least 15 contiguous nt of at least any one of SEQ ID NOS:1-37, preferably the entire sequence of at least any one of SEQ ID NOS:1-37, is not limited and can be any sequence of A, T, G, and/or C (for DNA) and A, U, G, and/or C (for RNA) or modified bases thereof, including inosine and pseudouridine. The choice of sequence will depend on the desired function and can be dictated by coding regions
- 20 desired, the intron-like regions desired, and the regulatory regions desired. Where the entire sequence of any one of SEQ ID NOS:1-37 is within the nucleic acid, the nucleic acid obtained is referred to herein as a polynucleotide comprising the sequence of any one of SEQ ID NOS:1-37.

Expression of Polypeptide Encoded by Full-Length cDNA or Full-Length Gene

- 25 The provided polynucleotides (e.g., a polynucleotide having a sequence of one of SEQ ID NOS:1-37), the corresponding cDNA, or the full-length gene is used to express a partial or complete gene product. Constructs of polynucleotides having sequences of SEQ ID NOS:1-37 can also be generated synthetically. Alternatively, single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides is described by, e.g.,
- 30 Stemmer *et al.*, *Gene (Amsterdam)* (1995) 164(1):49-53. In this method, assembly PCR (the synthesis of long DNA sequences from large numbers of oligodeoxyribonucleotides (oligos)) is described. The method is derived from DNA shuffling (Stemmer, *Nature* (1994) 370:389-

391), and does not rely on DNA ligase, but instead relies on DNA polymerase to build increasingly longer DNA fragments during the assembly process.

Appropriate polynucleotide constructs are purified using standard recombinant DNA techniques as described in, for example, Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual, 2nd Ed.*, (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY, and under current regulations described in United States Dept. of HHS, National Institute of Health (NIH) Guidelines for Recombinant DNA Research. The gene product encoded by a polynucleotide of the invention is expressed in any expression system, including, for example, bacterial, yeast, insect, amphibian and mammalian systems. Vectors, host cells and methods for obtaining expression in same are well known in the art. Suitable vectors and host cells are described in USPN 5,654,173.

Polynucleotide molecules comprising a polynucleotide sequence provided herein are generally propagated by placing the molecule in a vector. Viral and non-viral vectors are used, including plasmids. The choice of plasmid will depend on the type of cell in which propagation is desired and the purpose of propagation. Certain vectors are useful for amplifying and making large amounts of the desired DNA sequence. Other vectors are suitable for expression in cells in culture. Still other vectors are suitable for transfer and expression in cells in a whole animal or person. The choice of appropriate vector is well within the skill of the art. Many such vectors are available commercially. Methods for preparation of vectors comprising a desired sequence are well known in the art.

The polynucleotides set forth in SEQ ID NOS:1-37 or their corresponding full-length polynucleotides are linked to regulatory sequences as appropriate to obtain the desired expression properties. These can include promoters (attached either at the 5' end of the sense strand or at the 3' end of the antisense strand), enhancers, terminators, operators, repressors, and inducers. The promoters can be regulated or constitutive. In some situations it may be desirable to use conditionally active promoters, such as tissue-specific or developmental stage-specific promoters. These are linked to the desired nucleotide sequence using the techniques described above for linkage to vectors. Any techniques known in the art can be used.

When any of the above host cells, or other appropriate host cells or organisms, are used to replicate and/or express the polynucleotides or nucleic acids of the invention, the resulting replicated nucleic acid, RNA, expressed protein or polypeptide, is within the scope

of the invention as a product of the host cell or organism. The product is recovered by any appropriate means known in the art.

Once the gene corresponding to a selected polynucleotide is identified, its expression can be regulated in the cell to which the gene is native. For example, an endogenous gene of 5 a cell can be regulated by an exogenous regulatory sequence as disclosed in USPN 5,641,670. Identification of Functional and Structural Motifs of Novel Genes Screening Against Publicly Available Databases

Translations of the nucleotide sequence of the provided polynucleotides, cDNAs or full genes can be aligned with individual known sequences. Similarity with individual

10 sequences having a known activity can be used to determine the activity of the polypeptides encoded by the polynucleotides of the invention. Also, sequences exhibiting similarity with more than one individual sequence can exhibit activities that are characteristic of either or both individual sequences.

The full length sequences and fragments of the polynucleotide sequences of the 15 nearest neighbors, e.g., identified through BLAST searches using the provided polynucleotide sequences, can be used as probes and primers to identify and isolate the full length sequence corresponding to provided polynucleotides. The nearest neighbors can indicate a tissue or cell type to be used to construct a library for the full-length sequences corresponding to the provided polynucleotides.

20 Typically, a selected polynucleotide is translated in all six frames to determine the best alignment with the individual sequences. The sequences disclosed herein in the Sequence Listing are in a 5' to 3' orientation and translation in three frames can be sufficient (with a few specific exceptions as described in the Examples). These amino acid sequences are referred to, generally, as query sequences, which will be aligned with the individual 25 sequences. Databases with individual sequences are described in "Computer Methods for Macromolecular Sequence Analysis" *Methods in Enzymology* (1996) 266, Doolittle, Academic Press, Inc., a division of Harcourt Brace & Co., San Diego, California, USA. Databases include GenBank, EMBL, and DNA Database of Japan (DDBJ).

Query and individual sequences can be aligned using the methods and computer 30 programs described above, and include BLAST 2.0, available over the world wide web at <http://www.ncbi.nlm.nih.gov/BLAST/>. See also Altschul, et al. *Nucleic Acids Res.* (1997) 25:3389-3402. Another alignment algorithm is FASTA, available in the Genetics Computing Group (GCG) package, Madison, Wisconsin, USA, a wholly owned subsidiary of Oxford

Molecular Group, Inc. Other techniques for alignment are described in Doolittle, *supra*. Preferably, an alignment program that permits gaps in the sequence is utilized to align the sequences. The Smith-Waterman is one type of algorithm that permits gaps in sequence alignments. See *Meth. Mol. Biol.* (1997) 70: 173-187. Also, the GAP program using the 5 Needleman and Wunsch alignment method can be utilized to align sequences. An alternative search strategy uses MPSRCH software, which runs on a MASPAR computer. MPSRCH uses a Smith-Waterman algorithm to score sequences on a massively parallel computer. This approach improves ability to identify sequences that are distantly related matches, and is especially tolerant of small gaps and nucleotide sequence errors. Amino acid sequences 10 encoded by the provided polynucleotides can be used to search both protein and DNA databases. Incorporated herein by reference are all sequences that have been made public as of the filing date of this application by any of the DNA or protein sequence databases, including the patent databases (e.g., GeneSeq). Also incorporated by reference are those sequences that have been submitted to these databases as of the filing date of the present 15 application but not made public until after the filing date of the present application.

Results of individual and query sequence alignments can be divided into three categories: high similarity, weak similarity, and no similarity. Individual alignment results ranging from high similarity to weak similarity provide a basis for determining polypeptide activity and/or structure. Parameters for categorizing individual results include: percentage 20 of the alignment region length where the strongest alignment is found, percent sequence identity, and p value. The percentage of the alignment region length is calculated by counting the number of residues of the individual sequence found in the region of strongest alignment, *e.g.*, contiguous region of the individual sequence that contains the greatest number of residues that are identical to the residues of the corresponding region of the 25 aligned query sequence. This number is divided by the total residue length of the query sequence to calculate a percentage. For example, a query sequence of 20 amino acid residues might be aligned with a 20 amino acid region of an individual sequence. The individual sequence might be identical to amino acid residues 5, 9-15, and 17-19 of the query sequence. The region of strongest alignment is thus the region stretching from residue 9-19, an 11 30 amino acid stretch. The percentage of the alignment region length is: 11 (length of the region of strongest alignment) divided by (query sequence length) 20 or 55%.

Percent sequence identity is calculated by counting the number of amino acid matches between the query and individual sequence and dividing total number of matches by the

number of residues of the individual sequences found in the region of strongest alignment. Thus, the percent identity in the example above would be 10 matches divided by 11 amino acids, or approximately, 90.9%

P value is the probability that the alignment was produced by chance. For a single alignment, the p value can be calculated according to Karlin *et al.*, *Proc. Natl. Acad. Sci.* (1990) 87:2264 and Karlin *et al.*, *Proc. Natl. Acad. Sci.* (1993) 90. The p value of multiple alignments using the same query sequence can be calculated using an heuristic approach described in Altschul *et al.*, *Nat. Genet.* (1994) 6:119. Alignment programs such as BLAST program can calculate the p value. See also Altschul *et al.*, *Nucleic Acids Res.* (1997)

10 25.3389-3402.

Another factor to consider for determining identity or similarity is the location of the similarity or identity. Strong local alignment can indicate similarity even if the length of alignment is short. Sequence identity scattered throughout the length of the query sequence also can indicate a similarity between the query and profile sequences. The boundaries of the 15 region where the sequences align can be determined according to Doolittle, *supra*; BLAST 2.0 (see, e.g., Altschul, et al. *Nucleic Acids Res.* (1997) 25:3389-3402) or FAST programs; or by determining the area where sequence identity is highest.

High Similarity. In general, in alignment results considered to be of high similarity, the percent of the alignment region length is typically at least about 55% of total length query 20 sequence; more typically, at least about 58%; even more typically; at least about 60% of the total residue length of the query sequence. Usually, percent length of the alignment region can be as much as about 62%; more usually, as much as about 64%; even more usually, as much as about 66%. Further, for high similarity, the region of alignment, typically, exhibits at least about 75% of sequence identity; more typically, at least about 78%; even more 25 typically; at least about 80% sequence identity. Usually, percent sequence identity can be as much as about 82%; more usually, as much as about 84%; even more usually, as much as about 86%.

The p value is used in conjunction with these methods. If high similarity is found, the query sequence is considered to have high similarity with a profile sequence when the p value 30 is less than or equal to about 10^{-2} ; more usually; less than or equal to about 10^{-3} ; even more usually; less than or equal to about 10^{-4} . More typically, the p value is no more than about 10^{-5} ; more typically; no more than or equal to about 10^{-10} ; even more typically; no more than or equal to about 10^{-15} for the query sequence to be considered high similarity.

Weak Similarity. In general, where alignment results considered to be of weak similarity, there is no minimum percent length of the alignment region nor minimum length of alignment. A better showing of weak similarity is considered when the region of alignment is, typically, at least about 15 amino acid residues in length; more typically, at least 5 about 20; even more typically; at least about 25 amino acid residues in length. Usually, length of the alignment region can be as much as about 30 amino acid residues; more usually, as much as about 40; even more usually, as much as about 60 amino acid residues. Further, for weak similarity, the region of alignment, typically, exhibits at least about 35% of sequence identity; more typically, at least about 40%; even more typically; at least about 45% 10 sequence identity. Usually, percent sequence identity can be as much as about 50%; more usually, as much as about 55%; even more usually, as much as about 60%.

If low similarity is found, the query sequence is considered to have weak similarity with a profile sequence when the p value is usually less than or equal to about 10^{-2} ; more usually; less than or equal to about 10^{-3} ; even more usually; less than or equal to about 10^{-4} . 15 More typically, the p value is no more than about 10^{-5} ; more usually; no more than or equal to about 10^{-10} ; even more usually; no more than or equal to about 10^{-15} for the query sequence to be considered weak similarity.

Similarity Determined by Sequence Identity Alone. Sequence identity alone can be used to determine similarity of a query sequence to an individual sequence and can indicate 20 the activity of the sequence. Such an alignment, preferably, permits gaps to align sequences. Typically, the query sequence is related to the profile sequence if the sequence identity over the entire query sequence is at least about 15%; more typically, at least about 20%; even more typically, at least about 25%; even more typically, at least about 50%. Sequence identity alone as a measure of similarity is most useful when the query sequence is usually, at 25 least 80 residues in length; more usually, 90 residues; even more usually, at least 95 amino acid residues in length. More typically, similarity can be concluded based on sequence identity alone when the query sequence is preferably 100 residues in length; more preferably, 120 residues in length; even more preferably, 150 amino acid residues in length.

Alignments with Profile and Multiple Aligned Sequences. Translations of the 30 provided polynucleotides can be aligned with amino acid profiles that define either protein families or common motifs. Also, translations of the provided polynucleotides can be aligned to multiple sequence alignments (MSA) comprising the polypeptide sequences of members of protein families or motifs. Similarity or identity with profile sequences or MSAs can be used

to determine the activity of the gene products (*e.g.*, polypeptides) encoded by the provided polynucleotides or corresponding cDNA or genes. For example, sequences that show an identity or similarity with a chemokine profile or MSA can exhibit chemokine activities.

Profiles can be designed manually by (1) creating an MSA, which is an alignment of the

- 5 amino acid sequence of members that belong to the family and (2) constructing a statistical representation of the alignment. Such methods are described, for example, in Birney *et al.*, *Nucl. Acid Res.* (1996) 24(14): 2730-2739. MSAs of some protein families and motifs are publicly available. For example, <http://genome.wustl.edu/Pfam/> includes MSAs of 547 different families and motifs. These MSAs are described also in Sonnhammer *et al.*, *Proteins* 10 (1997) 28: 405-420. Other sources over the world wide web include the site at <http://www.embl-heidelberg.de/argos/ali/ali.html>; alternatively, a message can be sent to ALI@EMBL-HEIDELBERG.DE for the information. A brief description of these MSAs is reported in Pascarella *et al.*, *Prot. Eng.* (1996) 9(3):249-251. Techniques for building profiles from MSAs are described in Sonnhammer *et al.*, *supra*; Birney *et al.*, *supra*; and "Computer 15 Methods for Macromolecular Sequence Analysis," *Methods in Enzymology* (1996) 266, Doolittle, Academic Press, Inc., San Diego, California, USA.

Similarity between a query sequence and a protein family or motif can be determined by (a) comparing the query sequence against the profile and/or (b) aligning the query sequence with the members of the family or motif. Typically, a program such as Searchwise 20 is used to compare the query sequence to the statistical representation of the multiple alignment, also known as a profile (see Birney *et al.*, *supra*). Other techniques to compare the sequence and profile are described in Sonnhammer *et al.*, *supra* and Doolittle, *supra*.

Next, methods described by Feng *et al.*, *J. Mol. Evol.* (1987) 25:351 and Higgins *et al.*, *CABIOS* (1989) 5:151 can be used to align the query sequence with the members of a family 25 or motif, also known as a MSA. Sequence alignments can be generated using any of a variety of software tools. Examples include PileUp, which creates a multiple sequence alignment, and is described in Feng *et al.*, *J. Mol. Evol.* (1987) 25:351. Another method, GAP, uses the alignment method of Needleman *et al.*, *J. Mol. Biol.* (1970) 48:443. GAP is best suited for global alignment of sequences. A third method, BestFit, functions by inserting 30 gaps to maximize the number of matches using the local homology algorithm of Smith *et al.*, *Adv. Appl. Math.* (1981) 2:482. In general, the following factors are used to determine if a similarity between a query sequence and a profile or MSA exists: (1) number of conserved

residues found in the query sequence, (2) percentage of conserved residues found in the query sequence, (3) number of frameshifts, and (4) spacing between conserved residues.

- Some alignment programs that both translate and align sequences can make any number of frameshifts when translating the nucleotide sequence to produce the best
- 5 alignment. The fewer frameshifts needed to produce an alignment, the stronger the similarity or identity between the query and profile or MSAs. For example, a weak similarity resulting from no frameshifts can be a better indication of activity or structure of a query sequence, than a strong similarity resulting from two frameshifts. Preferably, three or fewer frameshifts are found in an alignment; more preferably two or fewer frameshifts; even more preferably,
- 10 one or fewer frameshifts; even more preferably, no frameshifts are found in an alignment of query and profile or MSAs.

- Conserved residues are those amino acids found at a particular position in all or some of the family or motif members. Alternatively, a position is considered conserved if only a certain class of amino acids is found in a particular position in all or some of the family
- 15 members. For example, the N-terminal position can contain a positively charged amino acid, such as lysine, arginine, or histidine.

- Typically, a residue of a polypeptide is conserved when a class of amino acids or a single amino acid is found at a particular position in at least about 40% of all class members; more typically, at least about 50%; even more typically, at least about 60% of the members.
- 20 Usually, a residue is conserved when a class or single amino acid is found in at least about 70% of the members of a family or motif; more usually, at least about 80%; even more usually, at least about 90%; even more usually, at least about 95%.

- A residue is considered conserved when three unrelated amino acids are found at a particular position in the some or all of the members; more usually, two unrelated amino
- 25 acids. These residues are conserved when the unrelated amino acids are found at particular positions in at least about 40% of all class member; more typically, at least about 50%; even more typically, at least about 60% of the members. Usually, a residue is conserved when a class or single amino acid is found in at least about 70% of the members of a family or motif; more usually, at least about 80%; even more usually, at least about 90%; even more usually,
- 30 at least about 95%.

A query sequence has similarity to a profile or MSA when the query sequence comprises at least about 25% of the conserved residues of the profile or MSA; more usually, at least about 30%; even more usually; at least about 40%. Typically, the query sequence has

a stronger similarity to a profile sequence or MSA when the query sequence comprises at least about 45% of the conserved residues of the profile or MSA; more typically, at least about 50%; even more typically; at least about 55%.

Identification of Secreted & Membrane-Bound Polypeptides

5 Both secreted and membrane-bound polypeptides of the present invention are of particular interest. For example, levels of secreted polypeptides can be assayed in body fluids that are convenient, such as blood, plasma, serum, and other body fluids such as urine, prostatic fluid and semen. Membrane-bound polypeptides are useful for constructing vaccine antigens or inducing an immune response. Such antigens would comprise all or part of the
10 extracellular region of the membrane-bound polypeptides. Because both secreted and membrane-bound polypeptides comprise a fragment of contiguous hydrophobic amino acids, hydrophobicity predicting algorithms can be used to identify such polypeptides.

A signal sequence is usually encoded by both secreted and membrane-bound polypeptide genes to direct a polypeptide to the surface of the cell. The signal sequence
15 usually comprises a stretch of hydrophobic residues. Such signal sequences can fold into helical structures. Membrane-bound polypeptides typically comprise at least one transmembrane region that possesses a stretch of hydrophobic amino acids that can transverse the membrane. Some transmembrane regions also exhibit a helical structure. Hydrophobic fragments within a polypeptide can be identified by using computer algorithms. Such
20 algorithms include Hopp & Woods, *Proc. Natl. Acad. Sci. USA* (1981) 78:3824-3828; Kyte & Doolittle, *J. Mol. Biol.* (1982) 157: 105-132; and RAOAR algorithm, Degli Esposti *et al.*, *Eur. J. Biochem.* (1990) 190: 207-219.

Another method of identifying secreted and membrane-bound polypeptides is to translate the polynucleotides of the invention in all six frames and determine if at least 8
25 contiguous hydrophobic amino acids are present. Those translated polypeptides with at least 8; more typically, 10; even more typically, 12 contiguous hydrophobic amino acids are considered to be either a putative secreted or membrane bound polypeptide. Hydrophobic amino acids include alanine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, threonine, tryptophan, tyrosine, and valine.

30 Identification of the Function of an Expression Product of a Full-Length Gene

Where the function of the encoded gene product is unknown, ribozymes, antisense constructs, and dominant negative mutants can be used to determine function of the expression product of a gene corresponding to a polynucleotide provided herein. These

methods and compositions are particularly useful where the provided novel polynucleotide exhibits no significant or substantial homology to a sequence encoding a gene of known function. Antisense molecules and ribozymes can be constructed from synthetic polynucleotides. Typically, the phosphoramidite method of oligonucleotide synthesis is used.

- 5 See Beaucage *et al.*, *Tet. Lett.* (1981) 22:1859 and USPN 4,668,777. Automated devices for synthesis are available to create oligonucleotides using this chemistry. Examples of such devices include Biosearch 8600, Models 392 and 394 by Applied Biosystems, a division of Perkin-Elmer Corp., Foster City, California, USA; and Expedite by Perceptive Biosystems, Framingham, Massachusetts, USA. Synthetic RNA, phosphate analog oligonucleotides, and
10 chemically derivatized oligonucleotides can also be produced, and can be covalently attached to other molecules. RNA oligonucleotides can be synthesized, for example, using RNA phosphoramidites. This method can be performed on an automated synthesizer, such as Applied Biosystems, Models 392 and 394, Foster City, California, USA.

Phosphorothioate oligonucleotides can also be synthesized for antisense construction.

- 15 A sulfurizing reagent, such as tetraethylthiuram disulfide (TETD) in acetonitrile can be used to convert the internucleotide cyanoethyl phosphite to the phosphorothioate triester within 15 minutes at room temperature. TETD replaces the iodine reagent, while all other reagents used for standard phosphoramidite chemistry remain the same. Such a synthesis method can be automated using Models 392 and 394 by Applied Biosystems, for example.
20 Oligonucleotides of up to 200 nt can be synthesized, more typically, 100 nt, more typically 50 nt; even more typically 30 to 40 nt. These synthetic fragments can be annealed and ligated together to construct larger fragments. See, for example, Sambrook *et al.*, *supra*. Trans-cleaving catalytic RNAs (ribozymes) are RNA molecules possessing
25 endoribonuclease activity. Ribozymes are specifically designed for a particular target, and the target message must contain a specific nucleotide sequence. They are engineered to cleave any RNA species site-specifically in the background of cellular RNA. The cleavage event renders the mRNA unstable and prevents protein expression. Importantly, ribozymes can be used to inhibit expression of a gene of unknown function for the purpose of determining its function in an in vitro or in vivo context, by detecting the phenotypic effect.
30 One commonly used ribozyme motif is the hammerhead, for which the substrate sequence requirements are minimal. Design of the hammerhead ribozyme, as well as therapeutic uses of ribozymes, are disclosed in Usman *et al.*, *Current Opin. Struct. Biol.* (1996) 6:527.

Methods for production of ribozymes, including hairpin structure ribozyme fragments, methods of increasing ribozyme specificity, and the like are known in the art.

The hybridizing region of the ribozyme can be modified or can be prepared as a branched structure as described in Horn and Urdea, *Nucleic Acids Res.* (1989) 17:6959. The

5 basic structure of the ribozymes can also be chemically altered in ways familiar to those skilled in the art, and chemically synthesized ribozymes can be administered as synthetic oligonucleotide derivatives modified by monomeric units. In a therapeutic context, liposome mediated delivery of ribozymes improves cellular uptake, as described in Birikh *et al.*, *Eur. J. Biochem.* (1997) 245:1.

10 Antisense nucleic acids are designed to specifically bind to RNA, resulting in the formation of RNA-DNA or RNA-RNA hybrids, with an arrest of DNA replication, reverse transcription or messenger RNA translation. Antisense polynucleotides based on a selected polynucleotide sequence can interfere with expression of the corresponding gene. Antisense polynucleotides are typically generated within the cell by expression from antisense
15 constructs that contain the antisense strand as the transcribed strand. Antisense polynucleotides based on the disclosed polynucleotides will bind and/or interfere with the translation of mRNA comprising a sequence complementary to the antisense polynucleotide. The expression products of control cells and cells treated with the antisense construct are compared to detect the protein product of the gene corresponding to the polynucleotide upon
20 which the antisense construct is based. The protein is isolated and identified using routine biochemical methods.

Given the extensive background literature and clinical experience in antisense therapy, one skilled in the art can use selected polynucleotides of the invention as additional potential therapeutics. The choice of polynucleotide can be narrowed by first testing them for
25 binding to "hot spot" regions of the genome of cancerous cells. If a polynucleotide is identified as binding to a "hot spot", testing the polynucleotide as an antisense compound in the corresponding cancer cells is warranted.

As an alternative method for identifying function of the gene corresponding to a polynucleotide disclosed herein, dominant negative mutations are readily generated for
30 corresponding proteins that are active as homomultimers. A mutant polypeptide will interact with wild-type polypeptides (made from the other allele) and form a non-functional multimer. Thus, a mutation is in a substrate-binding domain, a catalytic domain, or a cellular localization domain. Preferably, the mutant polypeptide will be overproduced. Point

mutations are made that have such an effect. In addition, fusion of different polypeptides of various lengths to the terminus of a protein can yield dominant negative mutants. General strategies are available for making dominant negative mutants (see, e.g., Herskowitz, *Nature* (1987) 329:219). Such techniques can be used to create loss of function mutations, which are 5 useful for determining protein function.

Polypeptides and Variants Thereof

Polypeptides contemplated by the present invention include those encoded by the disclosed polynucleotides and their corresponding full-length genes, as well as nucleic acids that, by virtue of the degeneracy of the genetic code, are not identical in sequence to the 10 disclosed polynucleotides. Thus, the invention includes within its scope a polypeptide encoded by a polynucleotide having the sequence of any one of SEQ ID NOS:1-37 or a variant thereof.

In general, the term "polypeptide" as used herein refers to both the full length polypeptide encoded by the recited polynucleotide, the polypeptide encoded by the gene 15 represented by the recited polynucleotide, as well as portions or fragments thereof (e.g., immunogenic fragments for production of specific antibodies, biologically active fragments that retain a biological activity of the native protein, etc.). "Polypeptides" also includes variants of the naturally occurring proteins, where such variants are homologous or substantially similar to the naturally occurring protein, and can be of an origin of the same or 20 different species as the naturally occurring protein (e.g., human, murine, or some other species that naturally expresses the recited polypeptide, usually a mammalian species). In general, variant polypeptides have a sequence that has at least about 80%, usually at least about 90%, and more usually at least about 98% sequence identity with a differentially expressed polypeptide of the invention, where amino acid sequence identity is determined 25 using the Smith-Waterman software program parameters described above. The variant polypeptides can be naturally or non-naturally glycosylated, i.e., the polypeptide has a glycosylation pattern that differs from the glycosylation pattern found in the corresponding naturally occurring protein.

The invention also encompasses homologs of the disclosed polypeptides (or 30 fragments thereof) where the homologs are isolated from other species, i.e. other animal or plant species, where such homologs, usually mammalian species, e.g. rodents, such as mice, rats; domestic animals, e.g., horse, cow, dog, cat; and humans. By "homolog" is meant a polypeptide having at least about 35%, usually at least about 40% and more usually at least

about 60% amino acid sequence identity to a particular differentially expressed protein as identified above.

In general, the polypeptides are provided in a non-naturally occurring environment, e.g. are separated from their naturally occurring environment. In certain embodiments, the 5 polypeptides are present in a composition that is enriched for the desired polypeptide as compared to a control. As such, purified polypeptide is provided, where by purified is meant that the protein is present in a composition that is substantially free of non-differentially expressed polypeptides, where by substantially free is meant that less than 90%, usually less than 60% and more usually less than 50% of the composition is made up of non-differentially 10 expressed polypeptides.

Also within the scope of the invention are variant polypeptides. Variants of polypeptides include mutants, fragments, and fusions. Mutants can include amino acid substitutions, additions or deletions. The amino acid substitutions can be conservative amino acid substitutions or substitutions to eliminate non-essential amino acids, such as to alter a 15 glycosylation site, a phosphorylation site or an acetylation site, or to minimize misfolding by substitution or deletion of one or more cysteine residues that are not necessary for function. Conservative amino acid substitutions are those that preserve the general charge, hydrophobicity/ hydrophilicity, and/or steric bulk of the amino acid substituted. Variants can be designed so as to retain or have enhanced biological activity of a particular region of the 20 protein (e.g., a functional domain and/or, where the polypeptide is a member of a protein family, a region associated with a consensus sequence). Selection of amino acid alterations for production of variants can be based upon the accessibility (interior vs. exterior) of the amino acid (see, e.g., Go *et al.*, *Int. J. Peptide Protein Res.* (1980) 15:211), the thermostability of the variant polypeptide (see, e.g., Querol *et al.*, *Prot. Eng.* (1996) 9:265), desired 25 glycosylation sites (see, e.g., Olsen and Thomsen, *J. Gen. Microbiol.* (1991) 137:579), desired disulfide bridges (see, e.g., Clarke *et al.*, *Biochemistry* (1993) 32:4322; and Wakarchuk *et al.*, *Protein Eng.* (1994) 7:1379), desired metal binding sites (see, e.g., Toma *et al.*, *Biochemistry* (1991) 30:97, and Haezerbrouck *et al.*, *Protein Eng.* (1993) 6:643), and desired substitutions within proline loops (see, e.g., Masul *et al.*, *Appl. Env. Microbiol.* 30 (1994) 60:3579). Cysteine-depleted muteins can be produced as disclosed in USPN 4,959,314.

Variants also include fragments of the polypeptides disclosed herein, particularly biologically active fragments and/or fragments corresponding to functional domains.

Fragments of interest will typically be at least about 10 aa to at least about 15 aa in length, usually at least about 50 aa in length, and can be as long as 300 aa in length or longer, but will usually not exceed about 1000 aa in length, where the fragment will have a stretch of amino acids that is identical to a polypeptide encoded by a polynucleotide having a sequence

- 5 of any SEQ ID NOS:1-37, or a homolog thereof. The protein variants described herein are encoded by polynucleotides that are within the scope of the invention. The genetic code can be used to select the appropriate codons to construct the corresponding variants.

Computer-Related Embodiments

In general, a library of polynucleotides is a collection of sequence information, which 10 information is provided in either biochemical form (*e.g.*, as a collection of polynucleotide molecules), or in electronic form (*e.g.*, as a collection of polynucleotide sequences stored in a computer-readable form, as in a computer system and/or as part of a computer program). The sequence information of the polynucleotides can be used in a variety of ways, *e.g.*, as a resource for gene discovery, as a representation of sequences expressed in a selected cell type 15 (*e.g.*, cell type markers), and/or as markers of a given disease or disease state. In general, a disease marker is a representation of a gene product that is present in all cells affected by disease either at an increased or decreased level relative to a normal cell (*e.g.*, a cell of the same or similar type that is not substantially affected by disease). For example, a polynucleotide sequence in a library can be a polynucleotide that represents an mRNA, 20 polypeptide, or other gene product encoded by the polynucleotide, that is either overexpressed or underexpressed in a breast ductal cell affected by cancer relative to a normal (*i.e.*, substantially disease-free) breast cell.

The nucleotide sequence information of the library can be embodied in any suitable form, *e.g.*, electronic or biochemical forms. For example, a library of sequence information 25 embodied in electronic form comprises an accessible computer data file (or, in biochemical form, a collection of nucleic acid molecules) that contains the representative nucleotide sequences of genes that are differentially expressed (*e.g.*, overexpressed or underexpressed) as between, for example, i) a cancerous cell and a normal cell; ii) a cancerous cell and a dysplastic cell; iii) a cancerous cell and a cell affected by a disease or condition other than 30 cancer; iv) a metastatic cancerous cell and a normal cell and/or non-metastatic cancerous cell; v) a malignant cancerous cell and a non-malignant cancerous cell (or a normal cell) and/or vi) a dysplastic cell relative to a normal cell. Other combinations and comparisons of cells affected by various diseases or stages of disease will be readily apparent to the ordinarily

skilled artisan. Biochemical embodiments of the library include a collection of nucleic acids that have the sequences of the genes in the library, where the nucleic acids can correspond to the entire gene in the library or to a fragment thereof, as described in greater detail below.

The polynucleotide libraries of the subject invention generally comprise sequence

5 information of a plurality of polynucleotide sequences, where at least one of the polynucleotides has a sequence of any of SEQ ID NOS:1-37. By plurality is meant at least 2, usually at least 3 and can include up to all of SEQ ID NOS:1-37. The length and number of polynucleotides in the library will vary with the nature of the library, *e.g.*, if the library is an oligonucleotide array, a cDNA array, a computer database of the sequence information, etc.

10 Where the library is an electronic library, the nucleic acid sequence information can be present in a variety of media. "Media" refers to a manufacture, other than an isolated nucleic acid molecule, that contains the sequence information of the present invention. Such a manufacture provides the genome sequence or a subset thereof in a form that can be examined by means not directly applicable to the sequence as it exists in a nucleic acid. For 15 example, the nucleotide sequence of the present invention, *e.g.* the nucleic acid sequences of any of the polynucleotides of SEQ ID NOS:1-37, can be recorded on computer readable media, *e.g.* any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as a floppy disc, a hard disc storage medium, and a magnetic tape; optical storage media such as CD-ROM; electrical 20 storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. One of skill in the art can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising a recording of the present sequence information. "Recorded" refers to a process for storing information on computer readable medium, using any such methods as known in the art. Any 25 convenient data storage structure can be chosen, based on the means used to access the stored information. A variety of data processor programs and formats can be used for storage, *e.g.* word processing text file, database format, *etc.* In addition to the sequence information, electronic versions of the libraries of the invention can be provided in conjunction or connection with other computer-readable information and/or other types of computer- 30 readable files (*e.g.*, searchable files, executable files, *etc.*, including, but not limited to, for example, search program software, *etc.*).

By providing the nucleotide sequence in computer readable form, the information can be accessed for a variety of purposes. Computer software to access sequence information is

publicly available. For example, the gapped BLAST (Altschul *et al.* *Nucleic Acids Res.* (1997) 25:3389-3402) and BLAZE (Brutlag *et al.* *Comp. Chem.* (1993) 17:203) search algorithms on a Sybase system can be used to identify open reading frames (ORFs) within the genome that contain homology to ORFs from other organisms.

5 As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available
10 computer-based system are suitable for use in the present invention. The data storage means can comprise any manufacture comprising a recording of the present sequence information as described above, or a memory access means that can access such a manufacture.

"Search means" refers to one or more programs implemented on the computer-based system, to compare a target sequence or target structural motif, or expression levels of a
15 polynucleotide in a sample, with the stored sequence information. Search means can be used to identify fragments or regions of the genome that match a particular target sequence or target motif. A variety of known algorithms are publicly known and commercially available, e.g. MacPattern (EMBL), BLASTN and BLASTX (NCBI). A "target sequence" can be any polynucleotide or amino acid sequence of six or more contiguous nucleotides or two or more
20 amino acids, preferably from about 10 to 100 amino acids or from about 30 to 300 nt. A variety of comparing means can be used to accomplish comparison of sequence information from a sample (e.g., to analyze target sequences, target motifs, or relative expression levels) with the data storage means. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the
25 computer based systems of the present invention to accomplish comparison of target sequences and motifs. Computer programs to analyze expression levels in a sample and in controls are also known in the art.

A "target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a
30 three-dimensional configuration that is formed upon the folding of the target motif, or on consensus sequences of regulatory or active sites. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzyme active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, hairpin structures,

promoter sequences and other expression elements such as binding sites for transcription factors.

A variety of structural formats for the input and output means can be used to input and output the information in the computer-based systems of the present invention. One format 5 for an output means ranks the relative expression levels of different polynucleotides. Such presentation provides a skilled artisan with a ranking of relative expression levels to determine a gene expression profile. .

As discussed above, the “library” of the invention also encompasses biochemical libraries of the polynucleotides of SEQ ID NOS:1-37, e.g., collections of nucleic acids 10 representing the provided polynucleotides. The biochemical libraries can take a variety of forms, e.g., a solution of cDNAs, a pattern of probe nucleic acids stably associated with a surface of a solid support (*i.e.*, an array) and the like. Of particular interest are nucleic acid arrays in which one or more of SEQ ID NOS:1-37 is represented on the array. By array is meant a an article of manufacture that has at least a substrate with at least two distinct nucleic 15 acid targets on one of its surfaces, where the number of distinct nucleic acids can be considerably higher, typically being at least 10 nt, usually at least 20 nt and often at least 25 nt. A variety of different array formats have been developed and are known to those of skill in the art. The arrays of the subject invention find use in a variety of applications, including gene expression analysis, drug screening, mutation analysis and the like, as disclosed in the 20 above-listed exemplary patent documents.

In addition to the above nucleic acid libraries, analogous libraries of polypeptides are also provided, where the where the polypeptides of the library will represent at least a portion of the polypeptides encoded by SEQ ID NOS:1-37.

Use of Polynucleotide Probes in Mapping, and in Tissue Profiling

25 Polynucleotide probes can be used for a variety of purposes, such as chromosome mapping of the polynucleotide and detection of transcription levels. Additional disclosure about preferred regions of the disclosed polynucleotide sequences is found in the Examples. A probe that hybridizes specifically to a polynucleotide disclosed herein should provide a detection signal at least 5-, 10-, or 20-fold higher than the background hybridization provided 30 with other unrelated sequences.

Detection of Expression Levels. Nucleotide probes are used to detect expression of a gene corresponding to the provided polynucleotide. In Northern blots, mRNA is separated electrophoretically and contacted with a probe. A probe is detected as hybridizing to an

mRNA species of a particular size. The amount of hybridization is quantitated to determine relative amounts of expression, for example under a particular condition. Probes are used for in situ hybridization to cells to detect expression. Probes can also be used *in vivo* for diagnostic detection of hybridizing sequences. Probes are typically labeled with a radioactive isotope. Other types of detectable labels can be used such as chromophores, fluors, and enzymes. Other examples of nucleotide hybridization assays are described in WO92/02526 and USPN 5,124,246.

Alternatively, the Polymerase Chain Reaction (PCR) is another means for detecting small amounts of target nucleic acids (see, e.g., Mullis *et al.*, *Meth. Enzymol.* (1987) 155:335; USPN 4,683,195; and USPN 4,683,202). Two primer polynucleotides nucleotides that hybridize with the target nucleic acids are used to prime the reaction: The primers can be composed of sequence within or 3' and 5' to the polynucleotides of the Sequence Listing. Alternatively, if the primers are 3' and 5' to these polynucleotides, they need not hybridize to them or the complements. After amplification of the target with a thermostable polymerase, the amplified target nucleic acids can be detected by methods known in the art, e.g., Southern blot. mRNA or cDNA can also be detected by traditional blotting techniques (e.g., Southern blot, Northern blot, etc.) described in Sambrook *et al.*, "Molecular Cloning: A Laboratory Manual" (New York, Cold Spring Harbor Laboratory, 1989) (e.g., without PCR amplification). In general, mRNA or cDNA generated from mRNA using a polymerase enzyme can be purified and separated using gel electrophoresis, and transferred to a solid support, such as nitrocellulose. The solid support is exposed to a labeled probe, washed to remove any unhybridized probe, and duplexes containing the labeled probe are detected.

Mapping. Polynucleotides of the present invention can be used to identify a chromosome on which the corresponding gene resides. Such mapping can be useful in identifying the function of the polynucleotide-related gene by its proximity to other genes with known function. Function can also be assigned to the polynucleotide-related gene when particular syndromes or diseases map to the same chromosome. For example, use of polynucleotide probes in identification and quantification of nucleic acid sequence aberrations is described in USPN 5,783,387. An exemplary mapping method is fluorescence in situ hybridization (FISH), which facilitates comparative genomic hybridization to allow total genome assessment of changes in relative copy number of DNA sequences (see, e.g., Valdes *et al.*, *Methods in Molecular Biology* (1997) 68:1). Polynucleotides can also be mapped to particular chromosomes using, for example, radiation hybrids or chromosome-

specific hybrid panels. See Leach *et al.*, *Advances in Genetics*, (1995) 33:63-99; Walter *et al.*, *Nature Genetics* (1994) 7:22; Walter and Goodfellow, *Trends in Genetics* (1992) 9:352. Panels for radiation hybrid mapping are available from Research Genetics, Inc., Huntsville, Alabama, USA. Databases for markers using various panels are available via the world wide web at <http://F/shgc-www.stanford.edu>; and <http://www-genome.wi.mit.edu/cgi-bin/contig/rhmapper.pl>. The statistical program RHMAP can be used to construct a map based on the data from radiation hybridization with a measure of the relative likelihood of one order versus another. RHMAP is available via the world wide web at <http://www.sph.umich.edu/group/statgen/software>. In addition, commercial programs are available for identifying regions of chromosomes commonly associated with disease, such as cancer.

Tissue Typing or Profiling. Expression of specific mRNA corresponding to the provided polynucleotides can vary in different cell types and can be tissue-specific. This variation of mRNA levels in different cell types can be exploited with nucleic acid probe assays to determine tissue types. For example, PCR, branched DNA probe assays, or blotting techniques utilizing nucleic acid probes substantially identical or complementary to polynucleotides listed in the Sequence Listing can determine the presence or absence of the corresponding cDNA or mRNA.

Tissue typing can be used to identify the developmental organ or tissue source of a metastatic lesion by identifying the expression of a particular marker of that organ or tissue. If a polynucleotide is expressed only in a specific tissue type, and a metastatic lesion is found to express that polynucleotide, then the developmental source of the lesion has been identified. Expression of a particular polynucleotide can be assayed by detection of either the corresponding mRNA or the protein product. As would be readily apparent to any forensic scientist, the sequences disclosed herein are useful in differentiating human tissue from non-human tissue. In particular, these sequences are useful to differentiate human tissue from bird, reptile, and amphibian tissue, for example.

Use of Polymorphisms. A polynucleotide of the invention can be used in forensics, genetic analysis, mapping, and diagnostic applications where the corresponding region of a gene is polymorphic in the human population. Any means for detecting a polymorphism in a gene can be used, including, but not limited to electrophoresis of protein polymorphic variants, differential sensitivity to restriction enzyme cleavage, and hybridization to allele-specific probes.

Antibody Production

Expression products of a polynucleotide of the invention, as well as the corresponding mRNA, cDNA, or complete gene, can be prepared and used for raising antibodies for experimental, diagnostic, and therapeutic purposes. For polynucleotides to which a

5 corresponding gene has not been assigned, this provides an additional method of identifying the corresponding gene. The polynucleotide or related cDNA is expressed as described above, and antibodies are prepared. These antibodies are specific to an epitope on the polypeptide encoded by the polynucleotide, and can precipitate or bind to the corresponding native protein in a cell or tissue preparation or in a cell-free extract of an in vitro expression

10 system.

Methods for production of antibodies that specifically bind a selected antigen are well known in the art. Immunogens for raising antibodies can be prepared by mixing a polypeptide encoded by a polynucleotide of the invention with an adjuvant, and/or by making fusion proteins with larger immunogenic proteins. Polypeptides can also be covalently linked to other larger immunogenic proteins, such as keyhole limpet hemocyanin. Immunogens are typically administered intradermally, subcutaneously, or intramuscularly to experimental animals such as rabbits, sheep, and mice, to generate antibodies. Monoclonal antibodies can be generated by isolating spleen cells and fusing myeloma cells to form hybridomas. Alternatively, the selected polynucleotide is administered directly, such as by intramuscular injection, and expressed in vivo. The expressed protein generates a variety of protein-specific immune responses, including production of antibodies, comparable to administration of the protein.

Preparations of polyclonal and monoclonal antibodies specific for polypeptides encoded by a selected polynucleotide are made using standard methods known in the art. The antibodies specifically bind to epitopes present in the polypeptides encoded by polynucleotides disclosed in the Sequence Listing. Typically, at least 6, 8, 10, or 12 contiguous amino acids are required to form an epitope. Epitopes that involve non-contiguous amino acids may require a longer polypeptide, e.g., at least 15, 25, or 50 amino acids. Antibodies that specifically bind to human polypeptides encoded by the provided polypeptides should provide a detection signal at least 5-, 10-, or 20-fold higher than a detection signal provided with other proteins when used in Western blots or other immunochemical assays. Preferably, antibodies that specifically polypeptides of the

invention do not bind to other proteins in immunochemical assays at detectable levels and can immunoprecipitate the specific polypeptide from solution.

The invention also contemplates naturally occurring antibodies specific for a polypeptide of the invention. For example, serum antibodies to a polypeptide of the

- 5 invention in a human population can be purified by methods well known in the art, e.g., by passing antiserum over a column to which the corresponding selected polypeptide or fusion protein is bound. The bound antibodies can then be eluted from the column, for example using a buffer with a high salt concentration.

- In addition to the antibodies discussed above, the invention also contemplates
10 genetically engineered antibodies, antibody derivatives (e.g., single chain antibodies, antibody fragments (e.g., Fab, etc.)), according to methods well known in the art.

Polynucleotides or Arrays for Diagnostics

Polynucleotide arrays provide a high throughput technique that can assay a large number of polynucleotide sequences in a sample. This technology can be used as a

- 15 diagnostic and as a tool to test for differential expression, e.g., to determine function of an encoded protein. Arrays can be created by spotting polynucleotide probes onto a substrate (e.g., glass, nitrocellulose, etc.) in a two-dimensional matrix or array having bound probes. The probes can be bound to the substrate by either covalent bonds or by non-specific interactions, such as hydrophobic interactions. Samples of polynucleotides can be detectably labeled (e.g.,
20 using radioactive or fluorescent labels) and then hybridized to the probes. Double stranded polynucleotides, comprising the labeled sample polynucleotides bound to probe polynucleotides, can be detected once the unbound portion of the sample is washed away.

- Techniques for constructing arrays and methods of using these arrays are described in EP 799 897; WO 97/29212; WO 97/27317; EP 785 280; WO 97/02357; USPN 5,593,839; USPN 25 5,578,832; EP 728 520; USPN 5,599,695; EP 721 016; USPN 5,556,752; WO 95/22058; and USPN 5,631,734. Arrays can be used to, for example, examine differential expression of genes and can be used to determine gene function. For example, arrays can be used to detect differential expression of a polynucleotide between a test cell and control cell (e.g., cancer cells and normal cells). For example, high expression of a particular message in a cancer
30 cell, which is not observed in a corresponding normal cell, can indicate a cancer specific gene product. Exemplary uses of arrays are further described in, for example, Pappalarado *et al.*, *Sem. Radiation Oncol.* (1998) 8:217; and Ramsay *Nature Biotechnol.* (1998) 16:40.

Differential Expression in Diagnosis

The polynucleotides of the invention can also be used to detect differences in expression levels between two cells, *e.g.*, as a method to identify abnormal or diseased tissue in a human. For polynucleotides corresponding to profiles of protein families, the choice of tissue can be selected according to the putative biological function. In general, the expression 5 of a gene corresponding to a specific polynucleotide is compared between a first tissue that is suspected of being diseased and a second, normal tissue of the human. The tissue suspected of being abnormal or diseased can be derived from a different tissue type of the human, but preferably it is derived from the same tissue type; for example an intestinal polyp or other abnormal growth should be compared with normal intestinal tissue. The normal tissue can be 10 the same tissue as that of the test sample, or any normal tissue of the patient, especially those that express the polynucleotide-related gene of interest (*e.g.*, brain, thymus, testis, heart, prostate, placenta, spleen, small intestine, skeletal muscle, pancreas, and the mucosal lining of the colon). A difference between the polynucleotide-related gene, mRNA, or protein in the two tissues which are compared, for example in molecular weight, amino acid or 15 nucleotide sequence, or relative abundance, indicates a change in the gene, or a gene which regulates it, in the tissue of the human that was suspected of being diseased. Examples of detection of differential expression and its use in diagnosis of cancer are described in USPNs 5,688,641 and 5,677,125.

A genetic predisposition to disease in a human can also be detected by comparing 20 expression levels of an mRNA or protein corresponding to a polynucleotide of the invention in a fetal tissue with levels associated in normal fetal tissue. Fetal tissues that are used for this purpose include, but are not limited to, amniotic fluid, chorionic villi, blood, and the blastomere of an in vitro-fertilized embryo. The comparable normal polynucleotide-related gene is obtained from any tissue. The mRNA or protein is obtained from a normal tissue of a 25 human in which the polynucleotide-related gene is expressed. Differences such as alterations in the nucleotide sequence or size of the same product of the fetal polynucleotide-related gene or mRNA, or alterations in the molecular weight, amino acid sequence, or relative abundance of fetal protein, can indicate a germline mutation in the polynucleotide-related gene of the fetus, which indicates a genetic predisposition to disease. In general, diagnostic, prognostic, 30 and other methods of the invention based on differential expression involve detection of a level or amount of a gene product, particularly a differentially expressed gene product, in a test sample obtained from a patient suspected of having or being susceptible to a disease (*e.g.*, breast cancer, lung cancer, colon cancer and/or metastatic forms thereof), and comparing the

detected levels to those levels found in normal cells (*e.g.*, cells substantially unaffected by cancer) and/or other control cells (*e.g.*, to differentiate a cancerous cell from a cell affected by dysplasia). Furthermore, the severity of the disease can be assessed by comparing the detected levels of a differentially expressed gene product with those levels detected in samples representing the levels of differentially gene product associated with varying degrees of severity of disease. It should be noted that use of the term "diagnostic" herein is not necessarily meant to exclude "prognostic" or "prognosis," but rather is used as a matter of convenience.

The term "differentially expressed gene" is generally intended to encompass a polynucleotide that can, for example, include an open reading frame encoding a gene product (*e.g.*, a polypeptide), and/or introns of such genes and adjacent 5' and 3' non-coding nucleotide sequences involved in the regulation of expression, up to about 20 kb beyond the coding region, but possibly further in either direction. The gene can be introduced into an appropriate vector for extrachromosomal maintenance or for integration into a host genome.

In general, a difference in expression level associated with a decrease in expression level of at least about 25%, usually at least about 50% to 75%, more usually at least about 90% or more is indicative of a differentially expressed gene of interest, *i.e.*, a gene that is underexpressed or down-regulated in the test sample relative to a control sample. Furthermore, a difference in expression level associated with an increase in expression of at least about 25%, usually at least about 50% to 75%, more usually at least about 90% and can be at least about 1 ½-fold, usually at least about 2-fold to about 10-fold, and can be about 100-fold to about 1,000-fold increase relative to a control sample is indicative of a differentially expressed gene of interest, *i.e.*, an overexpressed or up-regulated gene.

"Differentially expressed polynucleotide" as used herein means a nucleic acid molecule (RNA or DNA) comprising a sequence that represents a differentially expressed gene, *e.g.*, the differentially expressed polynucleotide comprises a sequence (*e.g.*, an open reading frame encoding a gene product) that uniquely identifies a differentially expressed gene so that detection of the differentially expressed polynucleotide in a sample is correlated with the presence of a differentially expressed gene in a sample. "Differentially expressed polynucleotides" is also meant to encompass fragments of the disclosed polynucleotides, *e.g.*, fragments retaining biological activity, as well as nucleic acids homologous, substantially similar, or substantially identical (*e.g.*, having about 90% sequence identity) to the disclosed polynucleotides.

"Diagnosis" as used herein generally includes determination of a subject's susceptibility to a disease or disorder, determination as to whether a subject is presently affected by a disease or disorder, as well as to the prognosis of a subject affected by a disease or disorder (*e.g.*, identification of pre-metastatic or metastatic cancerous states, stages of 5 cancer, or responsiveness of cancer to therapy). The present invention particularly encompasses diagnosis of subjects in the context of breast cancer (*e.g.*, carcinoma in situ (*e.g.*, ductal carcinoma in situ), estrogen receptor (ER)-positive breast cancer, ER-negative breast cancer, or other forms and/or stages of breast cancer), lung cancer (*e.g.*, small cell carcinoma, non-small cell carcinoma, mesothelioma, and other forms and/or stages of lung 10 cancer), and colon cancer (*e.g.*, adenomatous polyp, colorectal carcinoma, and other forms and/or stages of colon cancer).

"Sample" or "biological sample" as used throughout here are generally meant to refer to samples of biological fluids or tissues, particularly samples obtained from tissues, especially from cells of the type associated with the disease for which the diagnostic 15 application is designed (*e.g.*, ductal adenocarcinoma), and the like. "Samples" is also meant to encompass derivatives and fractions of such samples (*e.g.*, cell lysates). Where the sample is solid tissue, the cells of the tissue can be dissociated or tissue sections can be analyzed.

Methods of the subject invention useful in diagnosis or prognosis typically involve comparison of the abundance of a selected differentially expressed gene product in a sample 20 of interest with that of a control to determine any relative differences in the expression of the gene product, where the difference can be measured qualitatively and/or quantitatively. Quantitation can be accomplished, for example, by comparing the level of expression product detected in the sample with the amounts of product present in a standard curve. A 25 comparison can be made visually; by using a technique such as densitometry, with or without computerized assistance; by preparing a representative library of cDNA clones of mRNA isolated from a test sample, sequencing the clones in the library to determine that number of cDNA clones corresponding to the same gene product, and analyzing the number of clones corresponding to that same gene product relative to the number of clones of the same gene product in a control sample; or by using an array to detect relative levels of hybridization to a 30 selected sequence or set of sequences, and comparing the hybridization pattern to that of a control. The differences in expression are then correlated with the presence or absence of an abnormal expression pattern. A variety of different methods for determining the nucleic acid abundance in a sample are known to those of skill in the art (see, *e.g.*, WO 97/27317). In

general, diagnostic assays of the invention involve detection of a gene product of a the polynucleotide sequence (*e.g.*, mRNA or polypeptide) that corresponds to a sequence of SEQ ID NOS:1-1079. The patient from whom the sample is obtained can be apparently healthy, susceptible to disease (*e.g.*, as determined by family history or exposure to certain environmental factors), or can already be identified as having a condition in which altered expression of a gene product of the invention is implicated.

5 Diagnosis can be determined based on detected gene product expression levels of a gene product encoded by at least one, preferably at least two or more, at least 3 or more, or at least 4 or more of the polynucleotides having a sequence set forth in SEQ ID NOS:1-1079, 10 and can involve detection of expression of genes corresponding to all of SEQ ID NOS:1-1079 and/or additional sequences that can serve as additional diagnostic markers and/or reference sequences. Where the diagnostic method is designed to detect the presence or susceptibility of a patient to cancer, the assay preferably involves detection of a gene product encoded by a gene corresponding to a polynucleotide that is differentially expressed in cancer. Examples of 15 such differentially expressed polynucleotides are described in the Examples below. Given the provided polynucleotides and information regarding their relative expression levels provided herein, assays using such polynucleotides and detection of their expression levels in diagnosis and prognosis will be readily apparent to the ordinarily skilled artisan.

Any of a variety of detectable labels can be used in connection with the various 20 embodiments of the diagnostic methods of the invention. Suitable detectable labels include fluorochromes, (*e.g.* fluorescein isothiocyanate (FITC), rhodamine, Texas Red, phycoerythrin, allophycocyanin, 6-carboxyfluorescein (6-FAM), 2',7'-dimethoxy-4',5'-dichloro-6-carboxyfluorescein, 6-carboxy-X-rhodamine (ROX), 6-carboxy-2',4',7',4,7-hexachlorofluorescein (HEX), 5-carboxyfluorescein (5-FAM) or N,N,N',N'-tetramethyl-6-carboxyrhodamine (TAMRA)), radioactive labels, (*e.g.* ^{32}P , ^{35}S , ^3H , *etc.*), and the like. The detectable label can involve a two stage systems (*e.g.*, biotin-avidin, hapten-anti-hapten antibody, *etc.*)

Reagents specific for the polynucleotides and polypeptides of the invention, such as antibodies and nucleotide probes, can be supplied in a kit for detecting the presence of an 30 expression product in a biological sample. The kit can also contain buffers or labeling components, as well as instructions for using the reagents to detect and quantify expression products in the biological sample. Exemplary embodiments of the diagnostic methods of the invention are described below in more detail.

Polypeptide detection in diagnosis. In one embodiment, the test sample is assayed for the level of a differentially expressed polypeptide. Diagnosis can be accomplished using any of a number of methods to determine the absence or presence or altered amounts of the differentially expressed polypeptide in the test sample. For example, detection can utilize

5 staining of cells or histological sections with labeled antibodies, performed in accordance with conventional methods. Cells can be permeabilized to stain cytoplasmic molecules. In general, antibodies that specifically bind a differentially expressed polypeptide of the invention are added to a sample, and incubated for a period of time sufficient to allow binding to the epitope, usually at least about 10 minutes. The antibody can be detectably

10 labeled for direct detection (*e.g.*, using radioisotopes, enzymes, fluorescers, chemiluminescers, and the like), or can be used in conjunction with a second stage antibody or reagent to detect binding (*e.g.*, biotin with horseradish peroxidase-conjugated avidin, a secondary antibody conjugated to a fluorescent compound, *e.g.* fluorescein, rhodamine, Texas red, *etc.*). The absence or presence of antibody binding can be determined by various

15 methods, including flow cytometry of dissociated cells, microscopy, radiography, scintillation counting, *etc.* Any suitable alternative methods can of qualitative or quantitative detection of levels or amounts of differentially expressed polypeptide can be used, for example ELISA, western blot, immunoprecipitation, radioimmunoassay, etc.

mRNA detection. The diagnostic methods of the invention can also or alternatively

20 involve detection of mRNA encoded by a gene corresponding to a differentially expressed polynucleotides of the invention. Any suitable qualitative or quantitative methods known in the art for detecting specific mRNAs can be used. mRNA can be detected by, for example, *in situ* hybridization in tissue sections, by reverse transcriptase-PCR, or in Northern blots containing poly A+ mRNA. One of skill in the art can readily use these methods to

25 determine differences in the size or amount of mRNA transcripts between two samples.

mRNA expression levels in a sample can also be determined by generation of a library of expressed sequence tags (ESTs) from the sample, where the EST library is representative of sequences present in the sample (Adams, et al., (1991) *Science* 252:1651). Enumeration of the relative representation of ESTs within the library can be used to approximate the relative

30 representation of the gene transcript within the starting sample. The results of EST analysis of a test sample can then be compared to EST analysis of a reference sample to determine the relative expression levels of a selected polynucleotide, particularly a polynucleotide corresponding to one or more of the differentially expressed genes described herein.

Alternatively, gene expression in a test sample can be performed using serial analysis of gene expression (SAGE) methodology (e.g., Velculescu et al., *Science* (1995) 270:484) or differential display (DD) methodology (see, e.g., U.S. 5,776,683; and U.S. 5,807,680).

Alternatively, gene expression can be analyzed using hybridization analysis.

- 5 Oligonucleotides or cDNA can be used to selectively identify or capture DNA or RNA of specific sequence composition, and the amount of RNA or cDNA hybridized to a known capture sequence determined qualitatively or quantitatively, to provide information about the relative representation of a particular message within the pool of cellular messages in a sample. Hybridization analysis can be designed to allow for concurrent screening of the
- 10 relative expression of hundreds to thousands of genes by using, for example, array-based technologies having high density formats, including filters, microscope slides, or microchips, or solution-based technologies that use spectroscopic analysis (e.g., mass spectrometry). One exemplary use of arrays in the diagnostic methods of the invention is described below in more detail.
- 15 Use of a single gene in diagnostic applications. The diagnostic methods of the invention can focus on the expression of a single differentially expressed gene. For example, the diagnostic method can involve detecting a differentially expressed gene, or a polymorphism of such a gene (e.g., a polymorphism in an coding region or control region), that is associated with disease. Disease-associated polymorphisms can include deletion or
- 20 truncation of the gene, mutations that alter expression level and/or affect activity of the encoded protein, etc.

A number of methods are available for analyzing nucleic acids for the presence of a specific sequence, e.g. a disease associated polymorphism. Where large amounts of DNA are available, genomic DNA is used directly. Alternatively, the region of interest is cloned into a

25 suitable vector and grown in sufficient quantity for analysis. Cells that express a differentially expressed gene can be used as a source of mRNA, which can be assayed directly or reverse transcribed into cDNA for analysis. The nucleic acid can be amplified by conventional techniques, such as the polymerase chain reaction (PCR), to provide sufficient amounts for analysis, and a detectable label can be included in the amplification reaction

30 (e.g., using a detectably labeled primer or detectably labeled oligonucleotides) to facilitate detection. Alternatively, various methods are also known in the art that utilize oligonucleotide ligation as a means of detecting polymorphisms, see e.g., Riley et al., *Nucl. Acids Res.* (1990) 18:2887; and Delahunty et al., *Am. J. Hum. Genet.* (1996) 58:1239.

The amplified or cloned sample nucleic acid can be analyzed by one of a number of methods known in the art. The nucleic acid can be sequenced by dideoxy or other methods, and the sequence of bases compared to a selected sequence, *e.g.*, to a wild-type sequence. Hybridization with the polymorphic or variant sequence can also be used to determine its presence in a sample (*e.g.*, by Southern blot, dot blot, *etc.*). The hybridization pattern of a polymorphic or variant sequence and a control sequence to an array of oligonucleotide probes immobilized on a solid support, as described in US 5,445,934, or in WO 95/35505, can also be used as a means of identifying polymorphic or variant sequences associated with disease.

Single strand conformational polymorphism (SSCP) analysis, denaturing gradient gel electrophoresis (DGGE), and heteroduplex analysis in gel matrices are used to detect conformational changes created by DNA sequence variation as alterations in electrophoretic mobility. Alternatively, where a polymorphism creates or destroys a recognition site for a restriction endonuclease, the sample is digested with that endonuclease, and the products size fractionated to determine whether the fragment was digested. Fractionation is performed by gel or capillary electrophoresis, particularly acrylamide or agarose gels.

Screening for mutations in a gene can be based on the functional or antigenic characteristics of the protein. Protein truncation assays are useful in detecting deletions that can affect the biological activity of the protein. Various immunoassays designed to detect polymorphisms in proteins can be used in screening. Where many diverse genetic mutations lead to a particular disease phenotype, functional protein assays have proven to be effective screening tools. The activity of the encoded protein can be determined by comparison with the wild-type protein.

Pattern matching in diagnosis using arrays. In another embodiment, the diagnostic and/or prognostic methods of the invention involve detection of expression of a selected set of genes in a test sample to produce a test expression pattern (TEP). The TEP is compared to a reference expression pattern (REP), which is generated by detection of expression of the selected set of genes in a reference sample (*e.g.*, a positive or negative control sample). The selected set of genes includes at least one of the genes of the invention, which genes correspond to the polynucleotide sequences of SEQ ID NOS:1-1079. Of particular interest is a selected set of genes that includes gene differentially expressed in the disease for which the test sample is to be screened.

"Reference sequences" or "reference polynucleotides" as used herein in the context of differential gene expression analysis and diagnosis/prognosis refers to a selected set of

polynucleotides, which selected set includes at least one or more of the differentially expressed polynucleotides described herein. A plurality of reference sequences, preferably comprising positive and negative control sequences, can be included as reference sequences. Additional suitable reference sequences are found in GenBank, Unigene, and other nucleotide sequence databases (including, e.g., expressed sequence tag (EST), partial, and full-length sequences).

"Reference array" means an array having reference sequences for use in hybridization with a sample, where the reference sequences include all, at least one of, or any subset of the differentially expressed polynucleotides described herein. Usually such an array will include

- 5 at least 3 different reference sequences, and can include any one or all of the provided differentially expressed sequences. Arrays of interest can further comprise sequences, including polymorphisms, of other genetic sequences, particularly other sequences of interest for screening for a disease or disorder (e.g., cancer, dysplasia, or other related or unrelated diseases, disorders, or conditions). The oligonucleotide sequence on the array will usually be
- 10 at least about 12 nt in length, and can be of about the length of the provided sequences, or can extend into the flanking regions to generate fragments of 100 nt to 200 nt in length or more. Reference arrays can be produced according to any suitable methods known in the art. For example, methods of producing large arrays of oligonucleotides are described in U.S. 5,134,854, and U.S. 5,445,934 using light-directed synthesis techniques. Using a
- 15 computer controlled system, a heterogeneous array of monomers is converted, through simultaneous coupling at a number of reaction sites, into a heterogeneous array of polymers. Alternatively, microarrays are generated by deposition of pre-synthesized oligonucleotides onto a solid substrate, for example as described in PCT published application no. WO 95/35505.
- 20
- 25

- A "reference expression pattern" or "REP" as used herein refers to the relative levels of expression of a selected set of genes, particularly of differentially expressed genes, that is associated with a selected cell type, e.g., a normal cell, a cancerous cell, a cell exposed to an environmental stimulus, and the like. A "test expression pattern" or "TEP" refers to relative levels of expression of a selected set of genes, particularly of differentially expressed genes, in a test sample (e.g., a cell of unknown or suspected disease state, from which mRNA is isolated).
- 30

REPs can be generated in a variety of ways according to methods well known in the art. For example, REPs can be generated by hybridizing a control sample to an array having

a selected set of polynucleotides (particularly a selected set of differentially expressed polynucleotides), acquiring the hybridization data from the array, and storing the data in a format that allows for ready comparison of the REP with a TEP. Alternatively, all expressed sequences in a control sample can be isolated and sequenced, *e.g.*, by isolating mRNA from a

- 5 control sample, converting the mRNA into cDNA, and sequencing the cDNA. The resulting sequence information roughly or precisely reflects the identity and relative number of expressed sequences in the sample. The sequence information can then be stored in a format (*e.g.*, a computer-readable format) that allows for ready comparison of the REP with a TEP.

The REP can be normalized prior to or after data storage, and/or can be processed to

- 10 selectively remove sequences of expressed genes that are of less interest or that might complicate analysis (*e.g.*, some or all of the sequences associated with housekeeping genes can be eliminated from REP data).

TEPs can be generated in a manner similar to REPs, *e.g.*, by hybridizing a test sample to an array having a selected set of polynucleotides, particularly a selected set of

- 15 differentially expressed polynucleotides, acquiring the hybridization data from the array, and storing the data in a format that allows for ready comparison of the TEP with a REP. The REP and TEP to be used in a comparison can be generated simultaneously, or the TEP can be compared to previously generated and stored REPs.

In one embodiment of the invention, comparison of a TEP with a REP involves

- 20 hybridizing a test sample with a reference array, where the reference array has one or more reference sequences for use in hybridization with a sample. The reference sequences include all, at least one of, or any subset of the differentially expressed polynucleotides described herein. Hybridization data for the test sample is acquired, the data normalized, and the produced TEP compared with a REP generated using an array having the same or similar 25 selected set of differentially expressed polynucleotides. Probes that correspond to sequences differentially expressed between the two samples will show decreased or increased hybridization efficiency for one of the samples relative to the other.

Methods for collection of data from hybridization of samples with a reference arrays are well known in the art. For example, the polynucleotides of the reference and test samples 30 can be generated using a detectable fluorescent label, and hybridization of the polynucleotides in the samples detected by scanning the microarrays for the presence of the detectable label using, for example, a microscope and light source for directing light at a substrate. A photon counter detects fluorescence from the substrate, while an x-y translation

stage varies the location of the substrate. A confocal detection device that can be used in the subject methods is described in USPN 5,631,734. A scanning laser microscope is described in Shalon et al., *Genome Res.* (1996) 6:639. A scan, using the appropriate excitation line, is performed for each fluorophore used. The digital images generated from the scan are then
5 combined for subsequent analysis. For any particular array element, the ratio of the fluorescent signal from one sample (*e.g.*, a test sample) is compared to the fluorescent signal from another sample (*e.g.*, a reference sample), and the relative signal intensity determined.

Methods for analyzing the data collected from hybridization to arrays are well known in the art. For example, where detection of hybridization involves a fluorescent label, data
10 analysis can include the steps of determining fluorescent intensity as a function of substrate position from the data collected, removing outliers, *i.e.* data deviating from a predetermined statistical distribution, and calculating the relative binding affinity of the targets from the remaining data. The resulting data can be displayed as an image with the intensity in each region varying according to the binding affinity between targets and probes.

15 In general, the test sample is classified as having a gene expression profile corresponding to that associated with a disease or non-disease state by comparing the TEP generated from the test sample to one or more REPs generated from reference samples (*e.g.*, from samples associated with cancer or specific stages of cancer, dysplasia, samples affected by a disease other than cancer, normal samples, *etc.*). The criteria for a match or a substantial
20 match between a TEP and a REP include expression of the same or substantially the same set of reference genes, as well as expression of these reference genes at substantially the same levels (*e.g.*, no significant difference between the samples for a signal associated with a selected reference sequence after normalization of the samples, or at least no greater than about 25% to about 40% difference in signal strength for a given reference sequence. In
25 general, a pattern match between a TEP and a REP includes a match in expression, preferably a match in qualitative or quantitative expression level, of at least one of, all or any subset of the differentially expressed genes of the invention.

Pattern matching can be performed manually, or can be performed using a computer program. Methods for preparation of substrate matrices (*e.g.*, arrays), design of
30 oligonucleotides for use with such matrices, labeling of probes, hybridization conditions, scanning of hybridized matrices, and analysis of patterns generated, including comparison analysis, are described in, for example, U.S. 5,800,992.

Diagnosis, Prognosis and Management of Cancer

The polynucleotides of the invention and their gene products are of particular interest as genetic or biochemical markers (e.g., in blood or tissues) that will detect the earliest changes along the carcinogenesis pathway and/or to monitor the efficacy of various therapies 5 and preventive interventions. For example, the level of expression of certain polynucleotides can be indicative of a poorer prognosis, and therefore warrant more aggressive chemo- or radio-therapy for a patient or vice versa. The correlation of novel surrogate tumor specific features with response to treatment and outcome in patients can define prognostic indicators that allow the design of tailored therapy based on the molecular profile of the tumor. These 10 therapies include antibody targeting and gene therapy. Determining expression of certain polynucleotides and comparison of a patients profile with known expression in normal tissue and variants of the disease allows a determination of the best possible treatment for a patient, both in terms of specificity of treatment and in terms of comfort level of the patient. Surrogate tumor markers, such as polynucleotide expression, can also be used to better 15 classify, and thus diagnose and treat, different forms and disease states of cancer. Two classifications widely used in oncology that can benefit from identification of the expression levels of the polynucleotides of the invention are staging of the cancerous disorder, and grading the nature of the cancerous tissue.

The polynucleotides of the invention can be useful to monitor patients having or 20 susceptible to cancer to detect potentially malignant events at a molecular level before they are detectable at a gross morphological level. Furthermore, a polynucleotide of the invention identified as important for one type of cancer can also have implications for development or risk of development of other types of cancer, e.g., where a polynucleotide is differentially expressed across various cancer types. Thus, for example, expression of a polynucleotide 25 that has clinical implications for metastatic colon cancer can also have clinical implications for stomach cancer or endometrial cancer.

Staging. Staging is a process used by physicians to describe how advanced the cancerous state is in a patient. Staging assists the physician in determining a prognosis, planning treatment and evaluating the results of such treatment. Staging systems vary with 30 the types of cancer, but generally involve the following “TNM” system: the type of tumor, indicated by T; whether the cancer has metastasized to nearby lymph nodes, indicated by N; and whether the cancer has metastasized to more distant parts of the body, indicated by M. Generally, if a cancer is only detectable in the area of the primary lesion without having

spread to any lymph nodes it is called Stage I. If it has spread only to the closest lymph nodes, it is called Stage II. In Stage III, the cancer has generally spread to the lymph nodes in near proximity to the site of the primary lesion. Cancers that have spread to a distant part of the body, such as the liver, bone, brain or other site, are Stage IV, the most advanced stage.

- 5 The polynucleotides of the invention can facilitate fine-tuning of the staging process by identifying markers for the aggressivity of a cancer, e.g. the metastatic potential, as well as the presence in different areas of the body. Thus, a Stage II cancer with a polynucleotide signifying a high metastatic potential cancer can be used to change a borderline Stage II tumor to a Stage III tumor, justifying more aggressive therapy. Conversely, the presence of a
10 polynucleotide signifying a lower metastatic potential allows more conservative staging of a tumor.

Grading of cancers. Grade is a term used to describe how closely a tumor resembles normal tissue of its same type. The microscopic appearance of a tumor is used to identify tumor grade based on parameters such as cell morphology, cellular organization, and other
15 markers of differentiation. As a general rule, the grade of a tumor corresponds to its rate of growth or aggressiveness, with undifferentiated or high-grade tumors being more aggressive than well differentiated or low-grade tumors. The following guidelines are generally used for grading tumors: 1) GX Grade cannot be assessed; 2) G1 Well differentiated; G2 Moderately well differentiated; 3) G3 Poorly differentiated; 4) G4 Undifferentiated. The polynucleotides
20 of the invention can be especially valuable in determining the grade of the tumor, as they not only can aid in determining the differentiation status of the cells of a tumor, they can also identify factors other than differentiation that are valuable in determining the aggressiveness of a tumor, such as metastatic potential.

Detection of lung cancer. The polynucleotides of the invention can be used to detect
25 lung cancer in a subject. Although there are more than a dozen different kinds of lung cancer, the two main types of lung cancer are small cell and nonsmall cell, which encompass about 90% of all lung cancer cases. Small cell carcinoma (also called oat cell carcinoma) usually starts in one of the larger bronchial tubes, grows fairly rapidly, and is likely to be large by the time of diagnosis. Nonsmall cell lung cancer (NSCLC) is made up of three
30 general subtypes of lung cancer. Epidermoid carcinoma (also called squamous cell carcinoma) usually starts in one of the larger bronchial tubes and grows relatively slowly. The size of these tumors can range from very small to quite large. Adenocarcinoma starts growing near the outside surface of the lung and can vary in both size and growth rate. Some

slowly growing adenocarcinomas are described as alveolar cell cancer. Large cell carcinoma starts near the surface of the lung, grows rapidly, and the growth is usually fairly large when diagnosed. Other less common forms of lung cancer are carcinoid, cylindroma, mucoepidermoid, and malignant mesothelioma.

5 The polynucleotides of the invention, e.g., polynucleotides differentially expressed in normal cells versus cancerous lung cells (e.g., tumor cells of high or low metastatic potential) or between types of cancerous lung cells (e.g., high metastatic versus low metastatic), can be used to distinguish types of lung cancer as well as identifying traits specific to a certain patient's cancer and selecting an appropriate therapy. For example, if the patient's biopsy
10 expresses a polynucleotide that is associated with a low metastatic potential, it may justify leaving a larger portion of the patient's lung in surgery to remove the lesion. Alternatively, a smaller lesion with expression of a polynucleotide that is associated with high metastatic potential may justify a more radical removal of lung tissue and/or the surrounding lymph nodes, even if no metastasis can be identified through pathological examination.

15 Detection of breast cancer. The majority of breast cancers are adenocarcinomas subtypes, which can be summarized as follows: 1) ductal carcinoma in situ (DCIS), including comedocarcinoma; 2) infiltrating (or invasive) ductal carcinoma (IDC); 3) lobular carcinoma in situ (LCIS); 4) infiltrating (or invasive) lobular carcinoma (ILC); 5) inflammatory breast cancer; 6) medullary carcinoma; 7) mucinous carcinoma; 8) Paget's disease of the nipple; 9)
20 Phyllodes tumor; and 10) tubular carcinoma;

 The expression of polynucleotides of the invention can be used in the diagnosis and management of breast cancer, as well as to distinguish between types of breast cancer. Detection of breast cancer can be determined using expression levels of any of the appropriate polynucleotides of the invention, either alone or in combination. Determination
25 of the aggressive nature and/or the metastatic potential of a breast cancer can also be determined by comparing levels of one or more polynucleotides of the invention and comparing levels of another sequence known to vary in cancerous tissue, e.g. ER expression. In addition, development of breast cancer can be detected by examining the ratio of expression of a differentially expressed polynucleotide to the levels of steroid hormones (e.g.,
30 testosterone or estrogen) or to other hormones (e.g., growth hormone, insulin). Thus expression of specific marker polynucleotides can be used to discriminate between normal and cancerous breast tissue, to discriminate between breast cancers with different cells of origin, to discriminate between breast cancers with different potential metastatic rates, etc.

Detection of colon cancer. The polynucleotides of the invention exhibiting the appropriate expression pattern can be used to detect colon cancer in a subject. Colorectal cancer is one of the most common neoplasms in humans and perhaps the most frequent form of hereditary neoplasia. Prevention and early detection are key factors in controlling and

5 curing colorectal cancer. Colorectal cancer begins as polyps, which are small, benign growths of cells that form on the inner lining of the colon. Over a period of several years, some of these polyps accumulate additional mutations and become cancerous. Multiple familial colorectal cancer disorders have been identified, which are summarized as follows:

10 1) Familial adenomatous polyposis (FAP); 2) Gardner's syndrome; 3) Hereditary nonpolyposis colon cancer (HNPCC); and 4) Familial colorectal cancer in Ashkenazi Jews.

The expression of appropriate polynucleotides of the invention can be used in the diagnosis, prognosis and management of colorectal cancer. Detection of colon cancer can be determined using expression levels of any of these sequences alone or in combination with the levels of expression. Determination of the aggressive nature and/or the metastatic potential of a colon

15 cancer can be determined by comparing levels of one or more polynucleotides of the invention and comparing total levels of another sequence known to vary in cancerous tissue, e.g., expression of p53, DCC ras, or FAP (see, e.g., Fearon ER, et al., Cell (1990) 61(5):759; Hamilton SR et al., Cancer (1993) 72:957; Bodmer W, et al., Nat Genet. (1994) 4(3):217; Fearon ER, Ann N Y Acad Sci. (1995) 768:101). For example, development of colon cancer

20 can be detected by examining the ratio of any of the polynucleotides of the invention to the levels of oncogenes (e.g. ras) or tumor suppressor genes (e.g. FAP or p53). Thus expression of specific marker polynucleotides can be used to discriminate between normal and cancerous colon tissue, to discriminate between colon cancers with different cells of origin, to discriminate between colon cancers with different potential metastatic rates, etc.

25 Detection of prostate cancer. The polynucleotides and their corresponding genes and gene products exhibiting the appropriate differential expression pattern can be used to detect prostate cancer in a subject. Over 95% of primary prostate cancers are adenocarcinomas. Signs and symptoms may include: frequent urination, especially at night, inability to urinate, trouble starting or holding back urination, a weak or interrupted urine flow and frequent pain

30 or stiffness in the lower back, hips or upper thighs.

Many of the signs and symptoms of prostate cancer can be caused by a variety of other non-cancerous conditions. For example, one common cause of many of these signs and symptoms is a condition called benign prostatic hypertrophy, or BPH. In BPH, the prostate

gets bigger and may block the flow of urine or interfere with sexual function. The methods and compositions of the invention can be used to distinguish between prostate cancer and such non-cancerous conditions. The methods of the invention can be used in conjunction with conventional methods of diagnosis, e.g., digital rectal exam and/or detection of the level 5 of prostate specific antigen (PSA), a substance produced and secreted by the prostate.

Use of Polynucleotides to Screen for Peptide Analogs and Antagonists

Polypeptides encoded by the instant polynucleotides and corresponding full length genes can be used to screen peptide libraries to identify binding partners, such as receptors, from among the encoded polypeptides. Peptide libraries can be synthesized according to 10 methods known in the art (see, e.g., USPN 5,010,175 , and WO 91/17823). Agonists or antagonists of the polypeptides if the invention can be screened using any available method known in the art, such as signal transduction, antibody binding, receptor binding, mitogenic assays, chemotaxis assays, etc. The assay conditions ideally should resemble the conditions under which the native activity is exhibited *in vivo*, that is, under physiologic pH, 15 temperature, and ionic strength. Suitable agonists or antagonists will exhibit strong inhibition or enhancement of the native activity at concentrations that do not cause toxic side effects in the subject. Agonists or antagonists that compete for binding to the native polypeptide can require concentrations equal to or greater than the native concentration, while inhibitors capable of binding irreversibly to the polypeptide can be added in concentrations on the order 20 of the native concentration.

Such screening and experimentation can lead to identification of a novel polypeptide binding partner, such as a receptor, encoded by a gene or a cDNA corresponding to a polynucleotide of the invention, and at least one peptide agonist or antagonist of the novel binding partner. Such agonists and antagonists can be used to modulate, enhance, or inhibit 25 receptor function in cells to which the receptor is native, or in cells that possess the receptor as a result of genetic engineering. Further, if the novel receptor shares biologically important characteristics with a known receptor, information about agonist/antagonist binding can facilitate development of improved agonists/antagonists of the known receptor.

Pharmaceutical Compositions and Therapeutic Uses

30 Pharmaceutical compositions of the invention can comprise polypeptides, antibodies, or polynucleotides (including antisense nucleotides and ribozymes) of the claimed invention in a therapeutically effective amount. The term “therapeutically effective amount” as used herein refers to an amount of a therapeutic agent to treat, ameliorate, or prevent a desired

disease or condition, or to exhibit a detectable therapeutic or preventative (prophylactic) effect. The effect can be detected by, for example, chemical markers or antigen levels.

Therapeutic effects also include reduction in physical symptoms, such as decreased body temperature. The precise effective amount for a subject will depend upon the subject's size

- 5 and health, the nature and extent of the condition, and the therapeutics or combination of therapeutics selected for administration. Thus, it is not useful to specify an exact effective amount in advance. However, the effective amount for a given situation is determined by routine experimentation and is within the judgment of the clinician. For purposes of the present invention, an effective dose will generally be from about 0.01 mg/kg to 50 mg/kg or
10 0.05 mg/kg to about 10 mg/kg of the DNA constructs in the individual to which it is administered.

A pharmaceutical composition can also contain a pharmaceutically acceptable carrier.

The term "pharmaceutically acceptable carrier" refers to a carrier for administration of a therapeutic agent, such as antibodies or a polypeptide, genes, and other therapeutic agents.

- 15 The term refers to any pharmaceutical carrier that does not itself induce the production of antibodies harmful to the individual receiving the composition, and which can be administered without undue toxicity. Suitable carriers can be large, slowly metabolized macromolecules such as proteins, polysaccharides, polylactic acids, polyglycolic acids, polymeric amino acids, amino acid copolymers, and inactive virus particles. Such carriers
20 are well known to those of ordinary skill in the art. Pharmaceutically acceptable carriers in therapeutic compositions can include liquids such as water, saline, glycerol and ethanol. Auxiliary substances, such as wetting or emulsifying agents, pH buffering substances, and the like, can also be present in such vehicles. Typically, the therapeutic compositions are prepared as injectables, either as liquid solutions or suspensions; solid forms suitable for
25 solution in, or suspension in, liquid vehicles prior to injection can also be prepared. Liposomes are included within the definition of a pharmaceutically acceptable carrier. Pharmaceutically acceptable salts can also be present in the pharmaceutical composition, e.g., mineral acid salts such as hydrochlorides, hydrobromides, phosphates, sulfates, and the like; and the salts of organic acids such as acetates, propionates, malonates, benzoates, and the
30 like. A thorough discussion of pharmaceutically acceptable excipients is available in *Remington's Pharmaceutical Sciences* (Mack Pub. Co., N.J. 1991).

Delivery Methods. Once formulated, the compositions of the invention can be

- (1) administered directly to the subject (e.g., as polynucleotide or polypeptides); or (2)

delivered ex vivo, to cells derived from the subject (e.g., as in *ex vivo* gene therapy). Direct delivery of the compositions will generally be accomplished by parenteral injection, e.g., subcutaneously, intraperitoneally, intravenously or intramuscularly, intratumoral or to the interstitial space of a tissue. Other modes of administration include oral and pulmonary
5 administration, suppositories, and transdermal applications, needles, and gene guns or hyposprays. Dosage treatment can be a single dose schedule or a multiple dose schedule.

Methods for the *ex vivo* delivery and reimplantation of transformed cells into a subject are known in the art and described in e.g., International Publication No. WO 93/14778. Examples of cells useful in *ex vivo* applications include, for example, stem cells,
10 particularly hematopoietic, lymph cells, macrophages, dendritic cells, or tumor cells. Generally, delivery of nucleic acids for both *ex vivo* and *in vitro* applications can be accomplished by, for example, dextran-mediated transfection, calcium phosphate precipitation, polybrene mediated transfection, protoplast fusion, electroporation, encapsulation of the polynucleotide(s) in liposomes, and direct microinjection of the DNA
15 into nuclei, all well known in the art.

Once a gene corresponding to a polynucleotide of the invention has been found to correlate with a proliferative disorder, such as neoplasia, dysplasia, and hyperplasia, the disorder can be amenable to treatment by administration of a therapeutic agent based on the provided polynucleotide, corresponding polypeptide or other corresponding molecule (e.g.,
20 antisense, ribozyme, etc.).

The dose and the means of administration of the inventive pharmaceutical compositions are determined based on the specific qualities of the therapeutic composition, the condition, age, and weight of the patient, the progression of the disease, and other relevant factors. For example, administration of polynucleotide therapeutic compositions
25 agents of the invention includes local or systemic administration, including injection, oral administration, particle gun or catheterized administration, and topical administration. Preferably, the therapeutic polynucleotide composition contains an expression construct comprising a promoter operably linked to a polynucleotide of at least 12, 22, 25, 30, or 35 contiguous nt of the polynucleotide disclosed herein. Various methods can be used to
30 administer the therapeutic composition directly to a specific site in the body. For example, a small metastatic lesion is located and the therapeutic composition injected several times in several different locations within the body of tumor. Alternatively, arteries which serve a tumor are identified, and the therapeutic composition injected into such an artery, in order to

deliver the composition directly into the tumor. A tumor that has a necrotic center is aspirated and the composition injected directly into the now empty center of the tumor. The antisense composition is directly administered to the surface of the tumor, for example, by topical application of the composition. X-ray imaging is used to assist in certain of the above 5 delivery methods.

Receptor-mediated targeted delivery of therapeutic compositions containing an antisense polynucleotide, subgenomic polynucleotides, or antibodies to specific tissues can also be used. Receptor-mediated DNA delivery techniques are described in, for example, Findeis *et al.*, *Trends Biotechnol.* (1993) 11:202; Chiou *et al.*, *Gene Therapeutics: Methods 10 And Applications Of Direct Gene Transfer* (J.A. Wolff, ed.) (1994); Wu *et al.*, *J. Biol. Chem.* (1988) 263:621; Wu *et al.*, *J. Biol. Chem.* (1994) 269:542; Zenke *et al.*, *Proc. Natl. Acad. Sci. (USA)* (1990) 87:3655; Wu *et al.*, *J. Biol. Chem.* (1991) 266:338. Therapeutic compositions containing a polynucleotide are administered in a range of about 100 ng to about 200 mg of DNA for local administration in a gene therapy protocol. Concentration ranges of about 500 15 ng to about 50 mg, about 1 g to about 2 mg, about 5 g to about 500 g, and about 20 g to about 100 g of DNA can also be used during a gene therapy protocol. Factors such as method of action (e.g., for enhancing or inhibiting levels of the encoded gene product) and efficacy of transformation and expression are considerations which will affect the dosage required for ultimate efficacy of the antisense subgenomic polynucleotides. Where greater 20 expression is desired over a larger area of tissue, larger amounts of antisense subgenomic polynucleotides or the same amounts readministered in a successive protocol of administrations, or several administrations to different adjacent or close tissue portions of, for example, a tumor site, may be required to effect a positive therapeutic outcome. In all cases, routine experimentation in clinical trials will determine specific ranges for optimal 25 therapeutic effect. For polynucleotide related genes encoding polypeptides or proteins with anti-inflammatory activity, suitable use, doses, and administration are described in USPN 5,654,173.

The therapeutic polynucleotides and polypeptides of the present invention can be delivered using gene delivery vehicles. The gene delivery vehicle can be of viral or non-viral 30 origin (see generally, Jolly, *Cancer Gene Therapy* (1994) 1:51; Kimura, *Human Gene Therapy* (1994) 5:845; Connelly, *Human Gene Therapy* (1995) 1:185; and Kaplitt, *Nature Genetics* (1994) 6:148). Expression of such coding sequences can be induced using endogenous mammalian or heterologous promoters. Expression of the coding sequence can

be either constitutive or regulated.

Viral-based vectors for delivery of a desired polynucleotide and expression in a desired cell are well known in the art. Exemplary viral-based vehicles include, but are not limited to, recombinant retroviruses (see, e.g., WO 90/07936; WO 94/03622; WO 93/25698; 5 WO 93/25234; USPN 5,219,740; WO 93/11230; WO 93/10218; USPN 4,777,127; GB Patent No. 2,200,651; EP 0 345 242; and WO 91/02805), alphavirus-based vectors (e.g., Sindbis virus vectors, Semliki forest virus (ATCC VR-67; ATCC VR-1247), Ross River virus (ATCC VR-373; ATCC VR-1246) and Venezuelan equine encephalitis virus (ATCC VR-923; ATCC VR-1250; ATCC VR 1249; ATCC VR-532), and adeno-associated virus 10 (AAV) vectors (see, e.g., WO 94/12649, WO 93/03769; WO 93/19191; WO 94/28938; WO 95/11984 and WO 95/00655). Administration of DNA linked to killed adenovirus as described in Curiel, *Hum. Gene Ther.* (1992) 3:147 can also be employed.

Non-viral delivery vehicles and methods can also be employed, including, but not limited to, polycationic condensed DNA linked or unlinked to killed adenovirus alone (see, 15 e.g., Curiel, *Hum. Gene Ther.* (1992) 3:147); ligand-linked DNA (see, e.g., Wu, *J. Biol. Chem.* (1989) 264:16985); eukaryotic cell delivery vehicles cells (see, e.g., USPN 5,814,482; WO 95/07994; WO 96/17072; WO 95/30763; and WO 97/42338) and nucleic charge neutralization or fusion with cell membranes. Naked DNA can also be employed. Exemplary naked DNA introduction methods are described in WO 90/11092 and USPN 20 5,580,859. Liposomes that can act as gene delivery vehicles are described in USPN 5,422,120; WO 95/13796; WO 94/23697; WO 91/14445; and EP 0524968. Additional approaches are described in Philip, *Mol. Cell Biol.* (1994) 14:2411, and in Woffendin, *Proc. Natl. Acad. Sci.* (1994) 91:1581

Further non-viral delivery suitable for use includes mechanical delivery systems such 25 as the approach described in Woffendin *et al.*, *Proc. Natl. Acad. Sci. USA* (1994) 91(24):11581. Moreover, the coding sequence and the product of expression of such can be delivered through deposition of photopolymerized hydrogel materials or use of ionizing radiation (see, e.g., USPN 5,206,152 and WO 92/11033). Other conventional methods for gene delivery that can be used for delivery of the coding sequence include, for example, use 30 of hand-held gene transfer particle gun (see, e.g., USPN 5,149,655); use of ionizing radiation for activating transferred gene (see, e.g., USPN 5,206,152 and WO 92/11033).

The present invention will now be illustrated by reference to the following examples which set forth particularly advantageous embodiments. However, it should be noted that

these embodiments are illustrative and are not to be construed as restricting the invention in any way.

EXAMPLES

5 The present invention will now be illustrated by reference to the following examples which set forth particularly advantageous embodiments. However, it should be noted that these embodiments are illustrative and are not to be construed as restricting the invention in any way.

10 Example 1: Source of Biological Materials and Overview of Novel Polynucleotides
 Expressed by the Biological Materials

cDNA libraries were constructed from either human colon cancer cell line Km12L4-A (Morikawa, et al., *Cancer Research* (1988) 48:6863), KM12C (Morikawa et al. *Cancer Res.* (1988) 48:1943-1948), or MDA-MB-231 (Brinkley et al. *Cancer Res.* (1980) 40:3118-3129) 15 was used to construct a cDNA library from mRNA isolated from the cells. Sequences expressed by these cell lines were isolated and analyzed; most sequences were about 275-300 nucleotides in length. The KM12L4-A cell line is derived from the KM12C cell line. The KM12C cell line, which is poorly metastatic (low metastatic) was established in culture from a Dukes' stage B₂ surgical specimen (Morikawa et al. *Cancer Res.* (1988) 48:6863). The 20 KML4-A is a highly metastatic subline derived from KM12C (Yeatman et al. *Nucl. Acids. Res.* (1995) 23:4007; Bao-Ling et al. *Proc. Annu. Meet. Am. Assoc. Cancer. Res.* (1995) 21:3269). The KM12C and KM12C-derived cell lines (e.g., KM12L4, KM12L4-A, etc.) are well-recognized in the art as a model cell line for the study of colon cancer (see, e.g., Moriakawa et al., *supra*; Radinsky et al. *Clin. Cancer Res.* (1995) 1:19; Yeatman et al., 25 (1995) *supra*; Yeatman et al. *Clin. Exp. Metastasis* (1996) 14:246). The MDA-MB-231 cell line was originally isolated from pleural effusions (Cailleau, *J. Natl. Cancer. Inst.* (1974) 53:661), is of high metastatic potential, and forms poorly differentiated adenocarcinoma grade II in nude mice consistent with breast carcinoma.

30 Example 2: Differential Expression of Polynucleotides of the Invention: Description of
 Libraries and Detection of Differential Expression

The relative expression levels of various polynucleotides isolated from the Example 1 were assessed in several libraries prepared from various sources, including cell lines and

patient tissue samples. Table 1 provides a summary of these libraries, including the shortened library name (used hereafter), the mRNA source used to prepared the cDNA library, the "nickname" of the library that is used in the tables below (in quotes), and the approximate number of clones in the library.

5

Table 1. Description of cDNA Libraries

Library (lib #)	Description	No. of Clones in Library
1	Human Colon Cell Line Km12 L4: High Metastatic Potential (derived from Km12C)	308731
2	Human Colon Cell Line Km12C: Low Metastatic Potential	284771
3	Human Breast Cancer Cell Line MDA-MB-231: High Metastatic Potential; micro-metastases in lung	326937
4	Human Breast Cancer Cell Line MCF7: Non Metastatic	318979
8	Human Lung Cancer Cell Line MV-522: High Metastatic Potential	223620
9	Human Lung Cancer Cell Line UCP-3: Low Metastatic Potential	312503
12	Human microvascular endothelial cells (HMEC) - UNTREATED (PCR (OligodT) cDNA library)	41938
13	Human microvascular endothelial cells (HMEC) - bFGF TREATED (PCR (OligodT) cDNA library)	42100
14	Human microvascular endothelial cells (HMEC) - VEGF TREATED (PCR (OligodT) cDNA library)	42825
15	Normal Colon - UC#2 Patient (MICRODISSECTED PCR (OligodT) cDNA library)	282722
16	Colon Tumor - UC#2 Patient (MICRODISSECTED PCR (OligodT) cDNA library)	298831
17	Liver Metastasis from Colon Tumor of UC#2 Patient (MICRODISSECTED PCR (OligodT) cDNA library)	303467
18	Normal Colon - UC#3 Patient (MICRODISSECTED PCR (OligodT) cDNA library)	36216
19	Colon Tumor - UC#3 Patient (MICRODISSECTED PCR (OligodT) cDNA library)	41388
20	Liver Metastasis from Colon Tumor of UC#3 Patient (MICRODISSECTED PCR (OligodT) cDNA library)	30956
21	GRRpz Cells derived from normal prostate epithelium	164801
22	WOca Cells derived from Gleason Grade 4 prostate cancer epithelium	162088
23	Normal Lung Epithelium of Patient #1006 (MICRODISSECTED PCR (OligodT) cDNA library)	306198
24	Primary tumor, Large Cell Carcinoma of Patient #1006 (MICRODISSECTED PCR (OligodT) cDNA library)	309349

The KM12L4 and KM12C cell lines are described in Example 1 above. The MDA-MB-231 cell line was originally isolated from pleural effusions (Cailleau, *J. Natl. Cancer.*

10 *Inst.* (1974) 53:661), is of high metastatic potential, and forms poorly differentiated adenocarcinoma grade II in nude mice consistent with breast carcinoma. The MCF7 cell line was derived from a pleural effusion of a breast adenocarcinoma and is non-metastatic. The

MV-522 cell line is derived from a human lung carcinoma and is of high metastatic potential. The UCP-3 cell line is a low metastatic human lung carcinoma cell line; the MV-522 is a high metastatic variant of UCP-3. These cell lines are well-recognized in the art as models for the study of human breast and lung cancer (see, e.g., Chandrasekaran *et al.*, *Cancer Res.*

- 5 (1979) 39:870 (MDA-MB-231 and MCF-7); Gastpar *et al.*, *J Med Chem* (1998) 41:4965 (MDA-MB-231 and MCF-7); Ranson *et al.*, *Br J Cancer* (1998) 77:1586 (MDA-MB-231 and MCF-7); Kuang *et al.*, *Nucleic Acids Res* (1998) 26:1116 (MDA-MB-231 and MCF-7); Varki *et al.*, *Int J Cancer* (1987) 40:46 (UCP-3); Varki *et al.*, *Tumour Biol.* (1990) 11:327; (MV-522 and UCP-3); Varki *et al.*, *Anticancer Res.* (1990) 10:637; (MV-522); Kelner *et al.*,
10 *Anticancer Res* (1995) 15:867 (MV-522); and Zhang *et al.*, *Anticancer Drugs* (1997) 8:696 (MV522)). The samples of libraries 15-20 are derived from two different patients (UC#2, and UC#3). The bFGF-treated HMEC were prepared by incubation with bFGF at 10ng/ml for 2 hrs; the VEGF-treated HMEC were prepared by incubation with 20ng/ml VEGF for 2 hrs. Following incubation with the respective growth factor, the cells were washed and lysis
15 buffer added for RNA preparation. The GRRpz and WOca cell lines were provided by Dr. Donna M. Peehl, Department of Medicine, Stanford University School of Medicine. GRRpz was derived from normal prostate epithelium. The WOca cell line is a Gleason Grade 4 cell line.

- 20 Each of the libraries is composed of a collection of cDNA clones that in turn are representative of the mRNAs expressed in the indicated mRNA source. In order to facilitate the analysis of the millions of sequences in each library, the sequences were assigned to clusters. The concept of "cluster of clones" is derived from a sorting/grouping of cDNA clones based on their hybridization pattern to a panel of roughly 300 7bp oligonucleotide probes (see Drmanac *et al.*, *Genomics* (1996) 37(1):29). Random cDNA clones from a tissue
25 library are hybridized at moderate stringency to 300 7bp oligonucleotides. Each oligonucleotide has some measure of specific hybridization to that specific clone. The combination of 300 of these measures of hybridization for 300 probes equals the "hybridization signature" for a specific clone. Clones with similar sequence will have similar hybridization signatures. By developing a sorting/grouping algorithm to analyze these
30 signatures, groups of clones in a library can be identified and brought together computationally. These groups of clones are termed "clusters". Depending on the stringency of the selection in the algorithm (similar to the stringency of hybridization in a classic library cDNA screening protocol), the "purity" of each cluster can be controlled. For example,

artifacts of clustering may occur in computational clustering just as artifacts can occur in "wet-lab" screening of a cDNA library with 400 bp cDNA fragments, at even the highest stringency. The stringency used in the implementation of cluster herein provides groups of clones that are in general from the same cDNA or closely related cDNAs. Closely related 5 clones can be a result of different length clones of the same cDNA, closely related clones from highly related gene families, or splice variants of the same cDNA.

Differential expression for a selected cluster was assessed by first determining the number of cDNA clones corresponding to the selected cluster in the first library (Clones in 1st), and the determining the number of cDNA clones corresponding to the selected cluster in 10 the second library (Clones in 2nd). Differential expression of the selected cluster in the first library relative to the second library is expressed as a "ratio" of percent expression between the two libraries. In general, the "ratio" is calculated by: 1) calculating the percent expression of the selected cluster in the first library by dividing the number of clones corresponding to a selected cluster in the first library by the total number of clones analyzed from the first 15 library; 2) calculating the percent expression of the selected cluster in the second library by dividing the number of clones corresponding to a selected cluster in a second library by the total number of clones analyzed from the second library; 3) dividing the calculated percent expression from the first library by the calculated percent expression from the second library. If the "number of clones" corresponding to a selected cluster in a library is zero, the value is 20 set at 1 to aid in calculation. The formula used in calculating the ratio takes into account the "depth" of each of the libraries being compared, *i.e.*, the total number of clones analyzed in each library.

In general, a polynucleotide is said to be significantly differentially expressed between two samples when the ratio value is greater than at least about 2, preferably greater 25 than at least about 3, more preferably greater than at least about 5 , where the ratio value is calculated using the method described above. The significance of differential expression is determined using a z score test (Zar, Biostatistical Analysis, Prentice Hall, Inc., USA, "Differences between Proportions," pp 296-298 (1974).

Using the methods and libraries described above, 37 of the isolated polynucleotides 30 were identified as being differentially expressed across multiple libraries. Table 2 provides a list of these polynucleotides and their corresponding sequence names. The sequences of each of the above-referenced polynucleotides were determined using methods well known in the

art. The sequences of the 37 polynucleotides, assigned SEQ ID NOS:1-37, are provided in the Sequence Listing below.

Table 2 Polynucleotides corresponding to differentially expressed genes

SEQ ID NO.	Sequence Name	SEQ ID NO.	Sequence Name
1	13905	20	RTA00000683F.1.19.1
2	RTA00000281F.o.21.1	21	RTA00000172A.d.9.3
3	RTA00000348R.d.10.1	22	RTA00000165A.d.16.1
4	RTA00000177AF.d.22.3	23	RTA00000188AR.d.05.1
5	RTA00000684F.e.07.1	24	RTA00000183AF.n.14.1
6	RTA00000618F.p.24.1	25	RTA00000346F.g.11.1
7	RTA00000596F.d.12.1	26	RTA00000183AR.n.14.1
8	RTA00000421F.d.20.1	27	RTA00000742F.g.08.1
9	17090	28	RTA00000689F.h.06.1
10	RTA00000161A.1.7.1	29	RTA00000185AF.b.9.1
11	RTA00000155A.k.14.1	30	RTA0000018SAF.b.9.2
12	RTA00000163A.e.10.1	31	RTA00000192AR.o.8.2
13	RTA00000126A.o.15.2	32	RTA00000192AF.o.8.1
14	2546	33	RTA00000685F.j.16.1
15	RTA00000144A.p.8.1	34	RTA00000621F.i.13.2
16	RTA00000618F.k.16.1	35	RTA00000685F.1.23.1
17	RTA00000742F.o.19.1	36	16405
18	RTA00000148A.o.18.1	37	028035A
19	RTA00000619F.d.02.1		

The differential expression data for these sequences is provided below.

5 Example 3: Genes Differentially Expressed Genes in Non-Metastatic or Low Metastatic Potential Cancer Cells Versus High Metastatic Potential Cancer Cells

The relative levels of expression of genes corresponding to SEQ ID NO:1-37 across various libraries described in Table 1 are summarized in Table 3 below.

10 Table 3. Genes Differentially Expressed Across Multiple Library Comparisons

SEQ ID NO:	Cell or Tissue Sample and Cancer State Compared	RATIO
1	Low Met Breast (lib4) > High Met Breast (lib3)	5.38
1	Low Met Colon (lib2) > High Met Colon (lib1)	6.14
2	Low Met Colon (lib2) > High Met Colon (lib1)	3.56
2	Low Met Breast (lib4) > High Met Breast (lib3)	2.73
2	Normal Prostate (lib21) > Prostate Cancer (lib 22)	4.92
3	Low Met Colon (lib2) > High Met Colon (lib1)	3.52
3	Low Met Breast (lib4) > High Met Breast (lib3)	4.3
4	Low Met Colon (lib2) > High Met Colon (lib1)	3.52
4	Low Met Breast (lib4) > High Met Breast (lib3)	4.3
5	High Met Lung (lib8) > Low Met Lung (lib9)	3.35
5	Low Met Colon (lib2) > High Met Colon (lib1)	3.47

SEQ ID NO:	Cell or Tissue Sample and Cancer State Compared	RATIO
5	Low Met Breast (lib4) > High Met Breast (lib3)	30.24
6	Low Met Breast (lib4) > High Met Breast (lib3)	30.24
6	Low Met Colon (lib2) > High Met Colon (lib1)	3.47
6	High Met Lung (lib8) > Low Met Lung (lib9)	3.35
7	Low Met Colon (lib2) > High Met Colon (lib1)	3.47
7	Low Met Breast (lib4) > High Met Breast (lib3)	30.24
7	High Met Lung (lib8) > Low Met Lung (lib9)	3.35
8	Low Met Breast (lib4) > High Met Breast (lib3)	2.42
8	Low Met Colon (lib2) > High Met Colon (lib1)	2.63
9	Low Met Colon (lib2) > High Met Colon (lib1)	2.49
9	Low Met Breast (lib4) > High Met Breast (lib3)	2.19
9	Low Met Lung (lib9) > High Met Lung (lib8)	3.07
10	Low Met Breast (lib4) > High Met Breast (lib3)	41
10	High Met Lung (lib8) > Low Met Lung (lib9)	2.29
11	Low Met Breast (lib4) > High Met Breast (lib3)	7.35
11	Normal Prostate (lib21) > Prostate Cancer (lib 22)	9.84
12	High Met Breast (lib3) > Low Met Breast (lib4)	6.41
12	High Met Colon (lib1) > Low Met Colon (lib2)	2.39
13	High Met Colon (lib1) > Low Met Colon (lib2)	2.05
13	High Met Breast (lib3) > Low Met Breast (lib4)	9.76
14	Low Met Breast (lib4) > High Met Breast (lib3)	4.54
14	High Met Lung (lib8) > Low Met Lung (lib9)	10.48
14	Low Met Colon (lib2) > High Met Colon (lib1)	8.31
15	Low Met Breast (lib4) > High Met Breast (lib3)	2.05
15	Low Met Colon (lib2) > High Met Colon (lib1)	7.05
16	Low Met Colon (lib2) > High Met Colon (lib1)	4.34
16	Low Met Breast (lib4) > High Met Breast (lib3)	6.75
17	Low Met Colon (lib2) > High Met Colon (lib1)	4.34
17	Low Met Breast (lib4) > High Met Breast (lib3)	6.75
18	Low Met Colon (lib2) > High Met Colon (lib1)	3.98
18	Low Met Breast (lib4) > High Met Breast (lib3)	3.31
18	Low Met Lung (lib9) > High Met Lung (lib8)	2.5
19	Low Met Colon (lib2) > High Met Colon (lib1)	3.56
19	Normal Prostate (lib21) > Prostate Cancer (lib 22)	4.92
19	Low Met Breast (lib4) > High Met Breast (lib3)	2.73
20	Normal Prostate (lib21) > Prostate Cancer (lib 22)	4.92
20	Low Met Breast (lib4) > High Met Breast (lib3)	2.73
20	Low Met Colon (lib2) > High Met Colon (lib1)	3.56
21	Low Met Colon (lib2) > High Met Colon (lib1)	3.56
21	Low Met Breast (lib4) > High Met Breast (lib3)	2.73
21	Normal Prostate (lib21) > Prostate Cancer (lib 22)	4.92
22	Low Met Colon (lib2) > High Met Colon (lib1)	3.52
22	Low Met Breast (lib4) > High Met Breast (lib3)	3.55
22	High Met Lung (lib8) > Low Met Lung (lib9)	17.7
23	Low Met Colon (lib2) > High Met Colon (lib1)	3.25
23	Low Met Breast (lib4) > High Met Breast (lib3)	3.07
24	Low Met Breast (lib4) > High Met Breast (lib3)	3.07
24	Low Met Colon (lib2) > High Met Colon (lib1)	3.25

SEQ ID NO:	Cell or Tissue Sample and Cancer State Compared	RATIO
25	Low Met Colon (lib2) > High Met Colon (lib1)	3.25
25	Low Met Breast (lib4) > High Met Breast (lib3)	3.07
26	Low Met Colon (lib2) > High Met Colon (lib1)	3.25
26	Low Met Breast (lib4) > High Met Breast (lib3)	3.07
27	Low Met Colon (lib2) > High Met Colon (lib1)	3.25
27	Low Met Breast (lib4) > High Met Breast (lib3)	3.07
28	Low Met Colon (lib2) > High Met Colon (lib1)	2.86
28	Low Met Breast (lib4) > High Met Breast (lib3)	8.14
29	Low Met Colon (lib2) > High Met Colon (lib1)	2.1
29	Low Met Breast (lib4) > High Met Breast (lib3)	2.5
30	Low Met Colon (lib2) > High Met Colon (lib1)	2.1
30	Low Met Breast (lib4) > High Met Breast (lib3)	2.5
31	Low Met Colon (lib2) > High Met Colon (lib1)	2.1
31	Low Met Breast (lib4) > High Met Breast (lib3)	2.5
32	Low Met Colon (lib2) > High Met Colon (lib1)	2.1
32	Low Met Breast (lib4) > High Met Breast (lib3)	2.5
33	Low Met Colon (lib2) > High Met Colon (lib1)	2.14
33	Low Met Breast (lib4) > High Met Breast (lib3)	2.27
34	Normal Prostate (lib21) > Prostate Cancer (lib 22)	5.9
34	Low Met Colon (lib2) > High Met Colon (lib1)	2.1
34	Low Met Breast (lib4) > High Met Breast (lib3)	2.18
35	Normal Prostate (lib21) > Prostate Cancer (lib 22)	5.9
35	Low Met Colon (lib2) > High Met Colon (lib1)	2.1
35	Low Met Breast (lib4) > High Met Breast (lib3)	2.18
36	Low Met Colon (lib2) > High Met Colon (lib1)	2.1
36	Low Met Breast (lib4) > High Met Breast (lib3)	2.18
36	Normal Prostate (lib21) > Prostate Cancer (lib 22)	5.9
37	Low Met Colon (lib2) > High Met Colon (lib1)	2.17
37	Low Met Breast (lib4) > High Met Breast (lib3)	2.9
37	Low Met Lung (lib9) > High Met Lung (lib8)	3.4

Key for Table 5: High Met = high metastatic potential; Low Met = low metastatic potential; met = metastasized; tumor = non-metastasized tumor

The relative expression levels of the genes corresponding to the polynucleotides

- 5 above can be exploited in diagnostic and prognostic assays. For example, where the polynucleotide corresponds to a gene that is expressed at a relatively higher level in a low metastatic potential cell relative to a high metastatic potential cell (or at a relatively higher level in normal cells or nonmetastasized tumor cells relatively to metastatic or high metastatic potential cancerous cells), expression of the gene can serve as a marker indicating low risk of metastasis and may encode a suppressor of metastasis. Where the polynucleotide corresponds to a gene expressed at a relatively higher level in a high metastatic potential cell relative to a low metastatic potential cell, expression of the gene can serve as a marker of metastatic potential, indicating the need for more aggressive therapy.
- 10

Example 4: Identification of a Gene and Protein Encoded by the Polynucleotide

SEQ ID NOS:1-37 were translated in all three reading frames, and the nucleotide sequences and translated amino acid sequences used as query sequences to search for homologous sequences in either the GenBank (nucleotide sequences) or Non-Redundant Protein (amino acid sequences) databases. Query and individual sequences were aligned using the BLAST 2.0 programs, available over the world wide web at <http://www.ncbi.nlm.nih.gov/BLAST/>. (see also Altschul, et al. *Nucleic Acids Res.* (1997) 25:3389-3402). The sequences were masked to various extents to prevent searching of repetitive sequences or poly-A sequences, using the XBLAST program for masking low complexity.

The results are provided in Table 3 below.

15 **Table 4.** Results of search of publicly available sequence databases using SEQ ID NOS:1-37 as query sequences

SEQ ID NO:	Description
1	yt88d06.r1 Homo sapiens cDNA clone 231371 5'. (EST Accession No. H56522)
2	za04c10.r1 Soares melanocyte 2NbHM Homo sapiens cDNA clone 291570 5' (EST Accession No. W03386)
3	Homo sapiens heat shock factor binding protein 1 HSBP1 mRNA, complete cds (GenBank Accession No. AF068754)
4	Homo sapiens heat shock factor binding protein 1 HSBP1 mRNA, complete cds (GenBank Accession No. AF068754)
5	Homo sapiens CGI-122 protein mRNA, complete cds (GenBank Accession No. AF151880.1)
6	Homo sapiens CGI-122 protein mRNA, complete cds (GenBank Accession No. AF151880.1)
7	Homo sapiens CGI-122 protein mRNA, complete cds (GenBank Accession No. AF151880.1)
8	zn42b05.s1 Stratagene endothelial cell 937223 Homo sapiens cDNA clone 550065 3' similar to SW:RPC9_YEAST P28000 DNA-DIRECTED RNA POLYMERASES I AND III 16 KD POLYPEPTIDE (EST Accession No. AA102570)
9	yv31g09.r1 Soares fetal liver spleen 1NFLS Homo sapiens cDNA clone 244384 5' similar to contains Alu repetitive element (EST Accession No. N72329)
10	tz22h11.x1 NCI_CGAP_Ut2 Homo sapiens cDNA clone IMAGE:2289381 3', mRNA sequence (EST Accession No. AJ635233.1)
11	zi02h12.r1 Soares fetal liver spleen 1NFLS S1 Homo sapiens cDNA clone 429671 5' similar to contains Alu repetitive element (EST Accession No. AA011438)
12	Human quiescin (Q6) mRNA
13	Human Treacher Collins Syndrome
14	Human mRNA for annexin IV (carbohydrate-binding protein p33/41)
15	Human mRNA for TGIF protein
16	Human MHC class I lymphocyte antigen (HLA-E) (HLA-6.2)
17	Human HLA-E class I mRNA
18	Human Mpv17 mRNA

SEQ ID NO:	Description
19	Human kidney cyclophilin C
20	Human kidney cyclophilin C
21	Human kidney cyclophilin C
22	Human mRNA for 26S proteasome subunit p55
23	Human gamma-interferon-inducible protein (IP-30) mRNA
24	Human gamma-interferon-inducible protein (IP-30) mRNA
25	Human gamma-interferon-inducible protein (IP-30) mRNA
26	Human gamma-interferon-inducible protein (IP-30) mRNA
27	Human gamma-interferon-inducible protein (IP-30) mRNA
28	Human Na ⁺ /H ⁺ exchange regulatory co-factor (NHERF) mRNA
29	Human mRNA for mitochondrial dodecenoyl-CoA delta-isomerase
30	Human mRNA for mitochondrial dodecenoyl-CoA delta-isomerase
31	Human mRNA for mitochondrial dodecenoyl-CoA delta-isomerase
32	Human mRNA for mitochondrial dodecenoyl-CoA delta-isomerase
33	Human (clone PSK-J3) cyclin-dependent protein kinase mRNA
34	Human serine hydroxymethyltransferase mRNA
35	Human serine hydroxymethyltransferase mRNA
36	Human serine hydroxymethyltransferase mRNA
37	Human DNA damage-inducible RNA binding protein (A18hnRNP).

Key: ES = EST database; GB = GenBank database

SEQ ID NO:1 corresponds to a cDNA clone generated from an EST isolated from human pineal gland (Hillier *et al.* *Genome Res.* (1996) 6(9):807-28).

SEQ ID NO:2 corresponds to a sequence contained within a cDNA clone derived
5 from an EST isolated from a human melanocyte 2NbHM.

SEQ ID NOS:3 and 4 correspond to a sequence encoding a human heat shock factor binding protein, HSBP-1, which acts as a negative regulator of the heat shock response through its interaction with heat shock factor 1 (HSF1) (Satyal *et al.* *Genes Dev.* (1998) 12(13):1962-74). Briefly, HSF-1 responds to stress by undergoing conformational transition
10 from an inert non-DNA binding monomer to an active trimed that exhibits rapid DNA binding and activity as a transcriptional activator. Attenuation of the inducible transcriptional response, which occurs during heat shock or upon recovery at non-stress conditions, involves dissociation of the HSF1 trimer and loss of activity. HSBP-1, a nuclear-localized, conserved, 76-amino-acid protein, contains two extended arrays of hydrophobic repeats that interact with
15 HSF-1 heptad repeats of the active trimeric state of HSF1. During attenuation of HSF1 to the inert monomer, HSBP1 also associates with Hsp70. Through its interaction with HSF-1, HSBP1 negatively affects HSF-1 DNA-binding activity.

SEQ ID NOS:5-7 correspond to a gene encoding human CGI-122 protein.

SEQ ID NO:8 corresponds to a cDNA clone generated from an EST isolated from human endothelial cells (Hillier *et al.* *Genome Res.* (1996) 6(9):807-28).

SEQ ID NOS:9 and 11 correspond to a cDNA clone generated from an EST isolated from human fetal liver and spleen (Hillier *et al.* *Genome Res.* (1996) 6(9):807-28).

5 SEQ ID NO:10 corresponds to a sequence contained within a human cDNA clone isolated from moderately-differentiated endometrial adenocarcinoma.

The gene corresponding to SEQ ID NO:12 encodes human quiescin Q6 (Coppoch *et al.*, 1998, *Proc. Amer. Assoc. Can. Res.* 39:471).

The gene corresponding to SEQ ID NO:13 encodes a human Treacher Collins

10 Syndrome protein. Treacher Collins Syndrome (TCS) is an autosomal dominant disorder of craniofacial development including hearing loss and cleft palate. The TCS gene (called Treacle) has been positionally cloned and has 26 exons exhibiting a low complexity serine/alanine-rich protein of about 144 kDa (Dixon *et al.*, 1997, *Genome Res.* 7:223-234). Thirty-five mutations in the gene are reported from studies of individuals and families
15 affected by Treacher Collins Syndrome (Edwards *et al.*, 1997, *Am. J. Human Genet.* 60:515-524. Mutation in Treacle generally results in premature termination of the predicted protein (*Nat. Genet.* 12:130-136, 1996).

The gene corresponding to SEQ ID NO:14 encodes human annexin IV (carbohydrate-binding protein p33/41). Annexins are a family of Ca²⁺ and phospholipid binding proteins.
20 Annexin IV binds to glycosaminoglycans (GAGs) in a calcium-dependent manner (Kojima *et al.*, 1996, *J. Biol. Chem.* 271:7679-7685; Ishitsuka *et al.*, 1998, *J. Biol. Chem.* 273:9935-9941; and Satoh *et al.*, 1997, *Biol. Pharm. Bull.* 20:224-229). Annexin IV is highly expressed in various human adenocarcinoma cell lines (Satoh *et al.*, 1997, *FEBS Lett.* 405:107-110), and calcium-induced relocation of annexin IV is observed in a human osteosarcoma cell line
25 (Mohiti *et al.*, 1995, *Mol. Membr. Biol.* 12:321-329).

The gene corresponding to SEQ ID NO: 15 encodes human TGIF protein (Bertolino *et al.*, 1995, *J. Biol. Chem.* 270:31178-31188).

The gene corresponding to SEQ ID NO:16 encodes human MHC Class I lymphocyte antigen (HLA-E) (HLA-6.2), as described by Koller *et al.*, 1988, *J. Immunol.* 141:897-904.

30 The gene corresponding to SEQ ID NO:17 encodes human HLA-E class I mRNA, as described by Mizuno *et al.*, 1988, *J. Immunol.* 140:4024-4030.

The gene corresponding to SEQ ID NO:18 is the human glomerulosclerosis gene Mpv17, as described by Karasawa, 1993, *Hum. Mol. Genet.* 11:1829-1834.

The gene corresponding to any one or more of SEQ ID NOS:19-21 encodes a human cyclophilin C (Schneider *et al.*, 1994, *Biochemistry* 33:8218-8224).

The gene corresponding to SEQ ID NO:22 encodes human 26S proteasome subunit p55. Human 26S proteasome is a heterodimer of p44.5 and p55 (Saito *et al.*, 1997, *Gene* 203:241-250) and plays a major role in the non-lysosomal degradation of intracellular proteins (Mason *et al.*, 1998, *FEBS Lett.* 430:269-274). Homologues of 26S proteasome subunits are regulators of transcription and translation as described in Aravind and Ponting, 1998, *Protein Sci.* 7:1250-1254. Proteasomes are cylindrical particles made up of a stack of four heptameric rings (Rivett *et al.*, 1997, *Mol. Biol. Rep.* 24:99-102) and 26S proteasome has stringent organization of ATPases, as described in Seeger *et al.*, 1997, *Mol. Biol. Rep.* 24:83-88. In mammalian cells, the proteasome is a site for degradation of proteins, as described in Goldberg *et al.*, 1997, *Biol. Chem.* 378:131-140. In addition, proteolytic processing involving 26S proteasome occurs in lesions of Alzheimer's Disease and dementia with Lewy bodies (Fergusson *et al.*, 1996, *Neurosci. Lett.* 219:167-170).

The gene corresponding to any one or more of SEQ ID NOS:23-27 encodes human gamma-interferon-inducible protein (IP-30), Luster *et al.*, 1988, *J. Biol. Chem.* 263:12036-12043.

The gene corresponding to SEQ ID NO:28 encodes human Na⁺/H⁺ exchange regulatory co-factor (NHEFR) (Murphy *et al.*, 1998, *J. Biol. Chem.* in press).

The gene corresponding to any one or more of SEQ ID NOS:29-32 encodes human mitochondrial dodecenoyl-CoA delta-isomerase.

The gene corresponding to SEQ ID NO:33 encodes human (clone PSK-J3) cyclin-dependent protein kinase (Hanks, 1987, *Proc. Natl. Acad. Sci.* 84:388-392).

The gene corresponding to any one or more of SEQ ID NOS:34-36 encodes human serine hydroxymethyltransferase. Human serine hydroxymethyltransferase is a pyridoxine enzyme that is low in resting lymphocytes but increases upon antigenic or mitogenic stimuli, such as in an immune response (Trakatellis *et al.*, 1997, *Postgrad. Med. J.* 73:617-622, and Trakatellis *et al.*, 1994, *Postgrad. Med. J.* 70(Suppl 1):S89-S92). The catalytic function of the protein is tested as described in Kim *et al.*, 1997, *Anal. Biochem.* 253:201-209.

The polynucleotide comprising SEQ ID NO:37 corresponds to a GenBank entry having accession number AF021336, an mRNA complete coding sequence for human DNA damage-inducible RNA binding protein (A18hnRNP). The p value of 1.9⁻¹¹³ indicates an extremely high level of similarity between the sequence of SEQ ID NO:37 and the identified

GenBank sequence. Likewise, the protein search identified a high level of similarity (p value of 2.4^{-63}) between the amino acid translated from the second reading frame of the polynucleotide of SEQ ID NO:37 and the entry HUMCIRPA_1 for human mRNA for glycine-rich RNA binding protein cold-inducible RNA-binding protein (CIRP). The search 5 of DBEST identified accession number AA166551, murine CIRP, with a p value of 5.8^{-115} . CIRP is an 18kD protein induced in mouse cells by mild cold stress and consists of an N-terminal RNA-binding domain and a C-terminal glycine-rich domain (Nishiyama *et al.*, 1997, *J. Cell Biol.* 137(4):899). Lowering the culture temperature of BALB/3T3 cells from 37°C to 32°C induces CIRP expression and impairs cell growth. Suppression of CIRP with antisense 10 oligonucleotides alleviates the impaired growth, while overexpression of CIRP impairs growth at 37 °C and prolongs the G1 phase of the cell cycle (Nishiyama *et al.*, *supra*). The cloning and characterization of human CIRP was described by Nishiyama *et al.*, 1997, *Gene* 204(1-2):115.

15 **Deposit Information.** The materials described in Table 11 were deposited with the American Type Culture Collection (CMCC = Chiron Master Culture Collection).

Table 11. Cell Lines Deposited with ATCC

Cell Line	Deposit Date	ATCC Accession No.	CMCC Accession No.
KM12L4-A	March 19, 1998	CRL-12496	11606
Km12C	May 15, 1998	CRL-12533	11611
MDA-MB-231	May 15, 1998	CRL-12532	10583
MCF-7	October 9, 1998	CRL-12584	10377

20 The deposits described herein are provided merely as convenience to those of skill in the art, and is not an admission that a deposit is required under 35 U.S.C. §112. The sequence of the polynucleotides contained within the deposited material, as well as the amino acid sequence of the polypeptides encoded thereby, are incorporated herein by reference and are controlling in the event of any conflict with the written description of sequences herein.

25 A license may be required to make, use, or sell the deposited material, and no such license is granted hereby

Those skilled in the art will recognize, or be able to ascertain, using not more than routine experimentation, many equivalents to the specific embodiments of the invention

described herein. Such specific embodiments and equivalents are intended to be encompassed by the following claims.

All patents, published patent applications, and publications cited herein are incorporated by reference as if set forth fully herein.